Mixtures of exponential and geometric distributions, clumped Markov models with applications to biomedical research.

Tan, Jen Ning

Mixtures of Exponential and Geometric Distributions, Clumped Markov Models with Applications to Biomedical Research

by

Jen Ning Tan, BSc (Hons), University of Wales Swansea

Thesis submitted to the Swansea University

in candidature for the degree of

PHILOSOPHIÆ DOCTOR

School of Business and Economics
Swansea University
Singleton Park, Swansea SA2 8PP
United Kingdom

March 2010

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

JEN NING TAN
8 March 2010

Statement

This thesis is the result of my own investigation, except where acknowledgement of other sources is given.

JEN NING TAN
8 March 2010

Statement

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan (subject to the law of copyright), and for the title and summary to be made available to outside organisations.

JEN NING TAN
8 March 2010

Acknowledgements

This thesis would never have been possible without the endless support and encouragement from my principal supervisor, Dr. Assad Jalali. Thank you for introducing me to the world of mixture modelling. Your suggestions towards improving the quality of this thesis have always been stimulating and valuable. I also highly appreciate your permission to unveil your new estimation methods in this thesis. Above all, thank you for being so caring and understanding throughout these years.

I would like to express my heartfelt gratitude to my second supervisor, Professor Mike Gravenor, whose original ideas greatly enhanced the scope of this research and my understanding of prion disease. Thank you for your high level of support and encouragement.

I am truly grateful for the financial support from the ORS and the School of Business and Economics. A special thank you goes out to Professor Adriano Aguzzi and his laboratory team at the Institute of Neuropathology, University of Zurich, for kindly provided the experimental data for our study of prion disease. Many thanks also go to everyone in the School of Business and Economics for providing an enjoyable environment to work in.

A particular gratitude must be expressed to See Ju. A true friend reaches for your hand and touches your heart. I cannot imagine my PhD days without you. Another person I shall name individually is Jacky. Words cannot express how much I appreciate your care and support over the past few years. Finally, I would like to voice my heartfelt appreciation to my family for their love and care throughout my life.

JEN NING TAN

Swansea University
March 2010

This thesis was typeset with LaTeX 2_ε by the author. LaTeX 2_ε is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this thesis were written by Anton Merlushkin of the European Business Management School, Swansea University.

Summary

Finite mixture distributions are useful for modelling data with a range of different characteristics such as multi-modality, heterogeneity, skewness and kurtosis. Due to their flexibility, finite mixture models are widely applied in a diversity of areas such as medicine, genetics and marketing. However, the trade off for this flexibility is that the amount of algebra involved in estimation has created certain difficulties when using these models.

The first part of this thesis is focussed on the estimation problem for an exponential mixture, and its discrete analogue, a geometric mixture. We investigate a number of moment-based methods for estimating the parameters of a two-component exponential mixture distribution: (i) fractional moment estimator, (ii) attenuated moment estimator, (iii) Appell moment estimator, (iv) a method based on order statistics. For a mixture distribution, the traditional method of moments has long been out of favour due to its low efficiency compared to the more respectable maximum likelihood approach (MLE) via the Expectation-Maximisation (EM) algorithm. With a slight modification on the ordinary moments, we show, by our simulation experiments, that the efficiencies of the moment estimators are significantly improved. With a little amendment, these methods are then applied to the discrete analogue, a mixture of two geometric distributions. For both the continuous and discrete case, we observe high efficiencies in the parameter estimates given by the attenuated moment estimator, which is comparable to the MLE. With this method, users are able to obtain plausible estimates with just a calculator, or a spreadsheet. Compared to the MLE, the methods we investigate are undoubtedly quicker but at the same time provide reasonable parameter estimates, which might act as good starting points for the MLE via the EM algorithm.

In a hidden Markov process in which at least two states are clumped into a single level, unless the process is time reversible, the distribution of the sojourn time in the level is a linear combination of exponential distributions with at least one negative mixing weight. In this thesis, we investigate the performances of the aforementioned methods in estimating the parameters of a linear combination of two exponential distributions. We compare the efficiencies of these estimators to the asymptotically most efficient maximum likelihood estimator. We also study the application of these methods to the discrete analogue, a linear combination of two geometric distributions. The main purpose of our study is to provide users a quick but accurate method for the parameter estimation of such distributions. Our

investigations show that the MLE based on the EM algorithm is not an ideal method to identify a linear combination of distributions. On the other hand, the fractional moment estimator and the attenuated moment estimator return greatly improved parameter estimates over the MLE. Not to mention that these two methods are also quicker and do not require iterative process in order to obtain the parameter estimates.

In the second part of the thesis, we focus on developing new statistical models for the analysis of data from the prion diseases. Following from our theoretical investigations, we use mixture distributions to characterise the incubation period of chronic wasting disease in an experimental rodent system. We show that mixture models are a useful tool for capturing key features of the prion disease incubation period. Using statistical arguments alone, we are able to propose the presence in the experiment of two distinct prion strains, an observation that was confirmed by histopathology investigations. Using the same experimental framework (serial passage of prion disease) and a hidden markov process, we studied the probability that prion infection is transmitted sub-clinically to an exposed individual in experimental systems. A sub-clinically infected animal does not show clinical signs of scrapie but can potentially transmit it on to the next generation. Using an experimental system of scrapie disease in mice, we propose that the waiting time for a host to exhibit signs of scrapie can be modelled by a special kind of Markov process, Self Revealing Aggregated Markov Processes on Trees (SRAMPT). Our aim is to "track" the sub-clinically infected animals, giving an estimate of the overall prevalence of clinical scrapie in animals at each generation. In the serial passage study where sub-clinical infections cannot be detected experimentally, our model revealed a very strong involvement of sub-clinical status in the transmission of disease. We also show that de novo generated prions can generate sub-clinical infection at first passage at a much higher rate than previously assumed.

We then extend the prion sub-clinical model to allow application of the SRAMPT process to epidemic chains formed by contacts made during any infectious disease outbreak where sub-clinical infections may occur. The simple modification to the model greatly complicates the parameter estimation process. We solve the estimation problem by maximising the likelihood function according to SRAMPT and demonstrate its accuracy on a range of simulated examples. We show that the model is useful for estimating sub-clinical prevalence and transition rates under several scenarios.

Finally, we summarise our results and conclusions and discuss some directions for future research.

To my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Concepts, Methods and Tools

Finite mixture distributions have been recognised as useful statistical models due to their ability to capture unobserved heterogeneity in real data. With minimal effort, one can extend the traditional statistical model, using finite mixture distributions, to fit a variety of real data with different features, such as multimodality, skewness and kurtosis. Finite mixture models provide flexibility for statisticians when classical statistical models fail to fit real data. However, the trade off for such flexibility is that the computation of a mixture model is never an easy job. This explains why little work on mixture models had been done before the advent of computers. Pearson (1894) attempted the difficult task of estimating the five parameters of a mixture of two normal distributions in order to fit the measurements on the ratio of forehead to body length of crabs. Using the method of moments, the estimation problem was solved by calculating the roots of a nonic equation which is a polynomial equation of the ninth degree!

Ever since Dempster *et al.* (1977) introduced the Expectation-Maximisation algorithm, the method of moments has been disliked by most users due to its low relative efficiencies compared to the maximum likelihood estimator. The method of moments becomes the stepping stone for maximum likelihood estimator with the moment estimates being used as the starting values for the iterative procedure of the maximum likelihood estimator. With the advent of high speed computers, the maximum likelihood estimator becomes most users' favourite tool in mixture modelling because of its consistency and reliability. However, this method is time consuming and its computation is not as straightforward as the method of moments.

The work here is roughly divided into two sections. In the first part, several new methods constructed by Jalali (2005a, 2005b, 2005c and 2007) which have formal similarities to the method of moments are investigated in this thesis. The purpose is to provide an alternative for parameter estimation of finite mixture model, which is quick and easy to compute, with high efficiencies comparable to the maximum likelihood estimator. In particular, we

demonstrate how these methods are used in fitting mixtures of exponential distributions, and their discrete analogue, mixtures of geometric distributions. Both distributions are widely applicable in lifetime analysis.

The second part of this thesis presents the analysis of real biological data. We first use mixture models to describe the incubation period of a prion disease, scrapie, in serial passage experiments in mice. Lastly, we consider the problem of modelling the transmission probability of prion infection in these serial passage studies. A new model called Self Revealing Aggregated Markov Process on Tree (SRAMPT), constructed by Jalali (2008c), is used to estimate the sub-clinical infection on serial passage. We also describe use of the model as a potential tool for epidemiological analysis.

1.1 Statistical Distributions Studied

The main part of this thesis is about the study of mixtures of exponential distributions and its discrete analogue, mixtures of geometric distributions. Other famous distributions such as the Weibull distribution, Burr XII distribution, normal distribution, lognormal distribution and gamma distribution are also considered when modelling the heterogeneity in real lifetime data. Their basic properties are outlined here.

1.1.1 Exponential Distribution

The exponential distribution is widely used for modelling the lifetime of electronic components, for example light bulbs. This distribution is simple to use and is popular in reliability engineering. A continuous random variable T is said to have an exponential distribution with parameter θ if its probability density function (hereafter abbreviated to PDF) has the form

$$f(t;\theta) = \theta \exp(-\theta t), \tag{1.1}$$

where $t \geq 0$ and $\theta > 0$. An exponential distribution has higher probability for small t and lower probability for large t (see Figure 1.1). The cumulative distribution function (hereafter abbreviated to CDF) of an exponential distribution is given by

$$F(t;\theta) = 1 - \exp(-\theta t), \tag{1.2}$$

whereas the Laplace transform is

$$L(s) = \frac{\theta}{s+\theta}. \tag{1.3}$$

The expected value of the exponentially distributed random variable t is

$$\mu = E[T] = \frac{1}{\theta}, \tag{1.4}$$

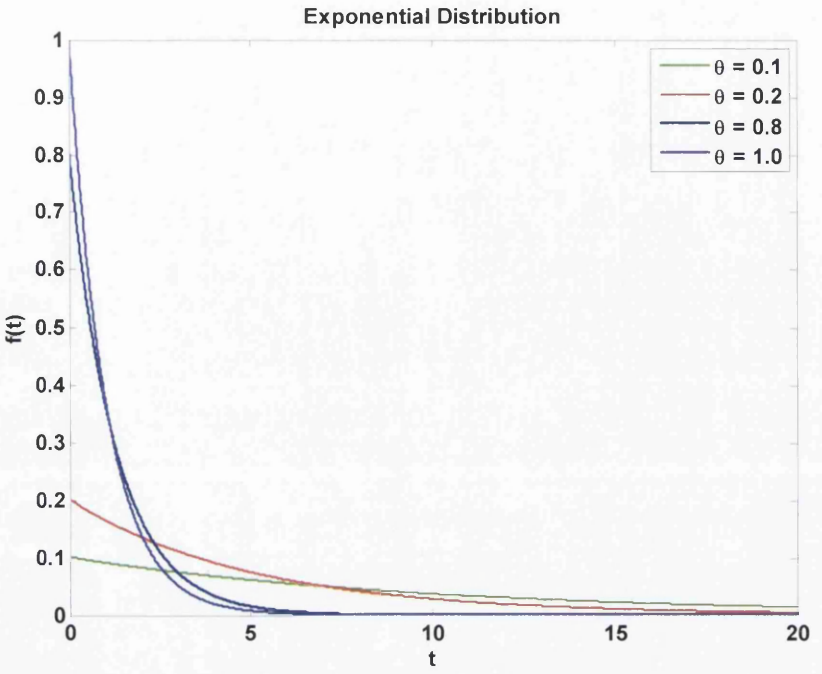


Figure 1.1: PDF plots of exponential distributions for varying θ.

and its variance is

$$\sigma^2 = Var[T] = \frac{1}{\theta^2}. \tag{1.5}$$

The median and mode for an exponential distribution are given by

$$\begin{aligned} \text{Median} &= \frac{\ln 2}{\theta} \\ \text{Mode} &= 0. \end{aligned} \tag{1.6}$$

The exponential distribution may be viewed as a continuous counterpart of the geometric distribution. The geometric distribution describes the number of independent Bernoulli trials necessary for a discrete process to change state. In contrast, the exponential distribution describes the time for a memoryless continuous process to change state.

An important property of the exponential distribution is memoryless. This means that no matter what has happened before, the probabilities associated with future performance, conditional upon past performance, will always be the same. This may be demonstrated mathematically as follows:

$$\Pr(T > s + t | T > t) = \Pr(T > s)$$

for all s, $t \geq 0$, i.e. conditional upon $T > t$, the probability that $T > t + s$ is the same as the unconditional probability that $T > s$.

For example, the lack of memory property says that the probability that a light bulb will function for at least one month, is no different from the conditional probability that the bulb will continue for a further month, given that it has already functioned for two weeks.

1.1.2 Geometric Distribution

The discrete analogue of the exponential distribution, the geometric distribution is frequently used to model the number of independent Bernoulli trials before the first success. For example, the number of tests required until a faulty electrical component is found. The geometric distribution is a discrete distribution with probability mass function (hereafter abbreviated to PMF)

$$f(n;\theta) = (1-\theta)^{n-1}\,\theta \tag{1.7}$$

for $n = 1, 2, 3, ...$where $0 \leq \theta \leq 1$ is the probability of getting one success. The CDF is given by

$$F(n;\theta) = 1 - (1-\theta)^{n}. \tag{1.8}$$

Figure 1.2 shows the PMF plots of geometric distributions for varying θ. The expected value of a geometrically distributed random variable N is

$$E[N] = \frac{1}{\theta} \tag{1.9}$$

and the variance is

$$Var[N] = \frac{1-\theta}{\theta^2}. \tag{1.10}$$

The probability-generating function of n is

$$G(s) = \frac{s\theta}{1-s(1-\theta)}. \tag{1.11}$$

The mode and median for a geometric distribution are given by

$$\begin{aligned} \text{Median} &= -\frac{\ln[2]}{\ln(1-\theta)} \\ \text{Mode} &= 1. \end{aligned} \tag{1.12}$$

Like the exponential distribution, the geometric distribution is memoryless. For example, if you roll a die until the first "1" appears, then the number of necessary additional trials is not affected by the fact that you have just observed a series without any "1s".

Note that the PMF for the geometric distribution can also be defined differently as

$$f(n;\theta) = (1-\theta)^{n}\,\theta \tag{1.13}$$

for $n = 0, 1, 2, ...$with expected value of N being $\frac{1-\theta}{\theta}$, and its variance is $\frac{1-\theta}{\theta^2}$. Throughout this thesis, we use the PMF form of the geometric distribution at (1.7) in most cases.

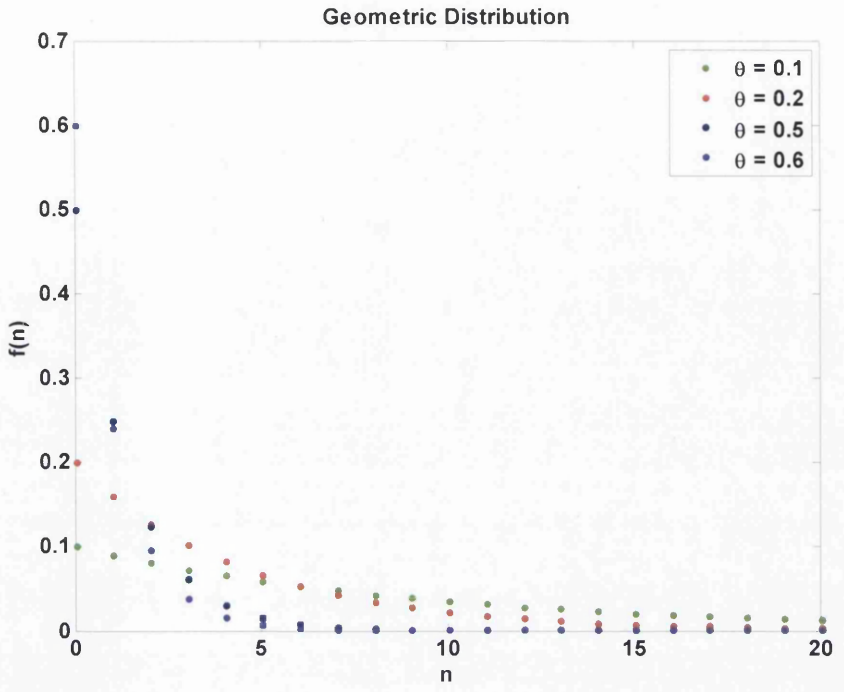


Figure 1.2: PMF plots of geometric distributions for varying θ.

1.1.3 Weibull Distribution

In reliability analysis, the Weibull distribution seems to be the favourite distribution of most statisticians for the fitting of survival data. This distribution, with a positive rate parameter $\theta > 0$ and a positive shape parameter $\alpha > 0$, has PDF

$$f(t;\theta,\alpha) = \alpha\theta^{\alpha}t^{\alpha-1}\exp\left[-(\theta t)^{\alpha}\right], \tag{1.14}$$

where $t \geq 0$ and CDF

$$F(t;\theta,\alpha) = 1 - \exp\left[-(\theta t)^{\alpha}\right]. \tag{1.15}$$

Both parameters are non-negative. The expected value of the Weibull distributed random variable t is

$$E[T] = \frac{1}{\theta}\Gamma\left(\frac{1}{\alpha}+1\right) \tag{1.16}$$

and its variance is

$$Var[T] = \frac{1}{\theta^2}\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \left[\Gamma\left(\frac{1}{\alpha}+1\right)\right]^2\right]. \tag{1.17}$$

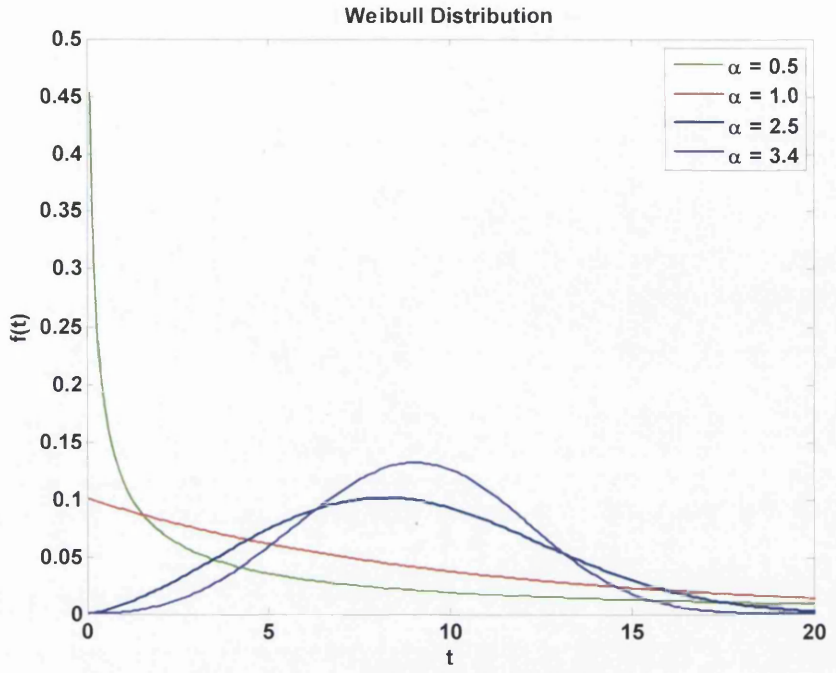


Figure 1.3: PDF plots of Weibull distributions for varying α; $\theta = 0.1$.

The median and mode of a Weibull distribution are given by

$$\begin{aligned} \text{Median} &= \frac{1}{\theta}\left(\ln[2]\right)^{\frac{1}{\alpha}} \\ \text{Mode} &= \frac{1}{\theta}\left(\frac{\alpha-1}{\alpha}\right)^{\frac{1}{\alpha}} \text{ if } \alpha > 1. \end{aligned} \tag{1.18}$$

This distribution, introduced by Weibull (1951), has a variety of shapes. Figure 1.3 shows how the Weibull PDF is affected by its shape parameter, α. When it is being used to analyse the lifetime data, a Weibull model suggests a decreasing failure rate if $\alpha < 1$. The exponential distribution is a special case of the Weibull distribution with $\alpha = 1$, where the failure rate is constant. If $\alpha > 1$, failure is more likely to occur as time goes on, i.e. increasing failure rate. A Weibull distribution behaves rather like a normal distribution when $\alpha = 3.4$. Due to its flexibility and ease of fitting, the Weibull distribution has been widely used as the lifetime distribution model.

1.1.4 Burr XII Distribution

The Burr XII distribution introduced by Burr (1942) has become increasingly popular in reliability analysis. The PDF for a random variable T following the three-parameter Burr XII distribution is

$$f(t;\alpha,\tau,\theta) = \frac{\tau\alpha\theta^{\tau}t^{\tau-1}}{(1+(\theta t)^{\tau})^{\alpha+1}} \tag{1.19}$$

for $t \geq 0$, with associated CDF

$$F(t;\alpha,\tau,\theta) = 1-(1+(\theta t)^{\tau})^{-\alpha} \tag{1.20}$$

in which both α and τ are positive shape parameters and θ is a positive rate parameter. The moments of T can be expressed as

$$E\left[T^k\right] = \frac{\Gamma\left(1+\frac{k}{\tau}\right)\Gamma\left(\alpha-\frac{k}{\tau}\right)}{\Gamma(\alpha)\,a^k}. \tag{1.21}$$

The theoretical mean and variance of a three parameter Burr XII distribution are given by

$$E[T] = \frac{1}{\tau\theta}B\left(\frac{1}{\tau},\alpha-\frac{1}{\tau}\right) \tag{1.22}$$

and

$$Var[T] = \frac{2}{\tau\theta^2}B\left(\frac{2}{\tau},\alpha-\frac{2}{\tau}\right)-\left[\frac{1}{\tau\theta}B\left(\frac{1}{\tau},\alpha-\frac{1}{\tau}\right)\right]^2. \tag{1.23}$$

It should be noted that a Weibull distribution is a special case of a Burr XII distribution (see Watkins (1999)), where the following relationship holds when $\frac{1}{\theta} \to \infty$ with $\alpha\theta$ remaining finite

$$F_B(t;\alpha,\tau,\theta) = F_W\left(t;\tau,\frac{1}{\theta\alpha^{\frac{1}{\tau}}}\right) \tag{1.24}$$

where F_B denotes the CDF of Burr XII distribution and F_W denotes the CDF of Weibull distribution.

The Burr XII distribution provides flexibility for modelling lifetime data because it can cover the curve shape characteristics for the normal, Weibull and gamma distribution, as shown in Figure 1.4.

1.1.5 Normal Distribution

The normal distribution is an important continuous distribution in a variety of fields. Let T be a random variable which is normally distributed with mean μ and variance $\sigma^2 > 0$, the PDF is given by

$$f(t;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] \tag{1.25}$$

whereas the CDF is

$$F(t;\mu,\sigma) = \frac{1}{2}\left[1+\operatorname{erf}\left[\frac{t-\mu}{\sigma\sqrt{2}}\right]\right], \tag{1.26}$$

where $\operatorname{erf}[x]$ is the Gauss error function, defined as

$$\operatorname{erf}[x] = \frac{2}{\sqrt{\pi}}\int_0^x \exp\left[-t^2\right]dt. \tag{1.27}$$

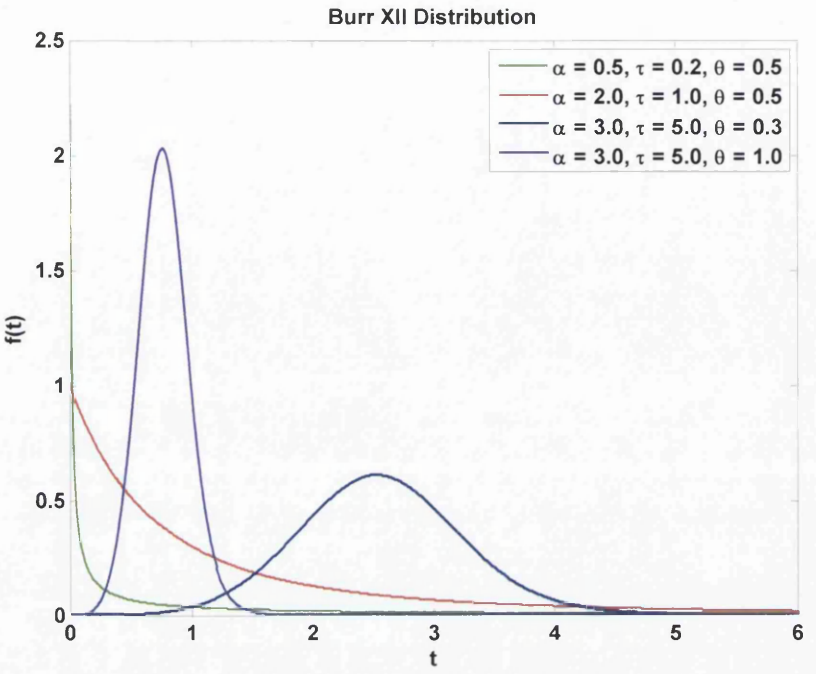


Figure 1.4: PDF plots of Burr XII distributions for varying α, τ and θ.

The PDF is bell shaped and is symmetrical about its mean μ. The inflexion points of the curve occur at one standard deviation σ away from the mean. For a normal distribution, the mean, mode and median are given by μ; while the variance is σ^2. Figure 1.5 shows the PDF for various values of μ and σ.

1.1.6 Lognormal Model

A positive random variable T is said to have a lognormal distribution in case $\ln T$ has a normal distribution. Clearly such a random variable has a continuous distribution and its associated PDF is defined as

$$f\left(t;\mu,\sigma\right)=\frac{1}{t\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(\ln\left[t\right]-\mu\right)^{2}}{2\sigma^{2}}\right], \tag{1.28}$$

and CDF

$$F\left(t;\mu,\sigma\right)=\frac{1}{2}+\frac{1}{2}\operatorname{erf}\left[\frac{\ln\left[t\right]-\mu}{\sigma\sqrt{2}}\right] \tag{1.29}$$

where $t>0$, $-\infty<\mu<\infty$ and $\sigma>0$. If T is a lognormally distributed variable, its expected value is

$$E\left[T\right]=\exp\left[\mu+\frac{\sigma^{2}}{2}\right] \tag{1.30}$$

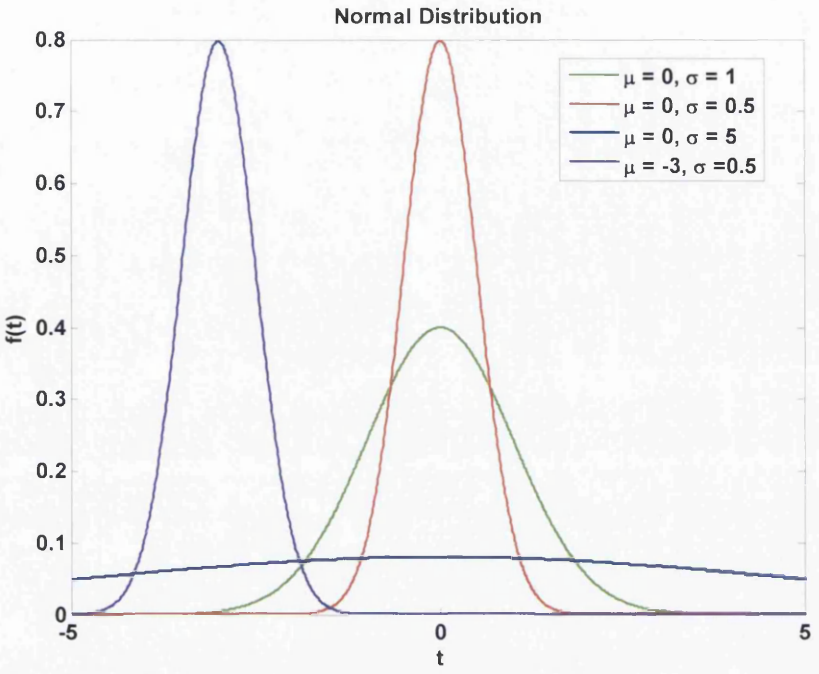


Figure 1.5: PDF plots of normal distributions for varying μ and σ.

and its variance is

$$Var[T] = \left(\exp\left[\sigma^2\right] - 1\right)\exp\left[2\mu + \sigma^2\right]. \tag{1.31}$$

The median and mode of T are defined as

$$\begin{aligned} \text{Median} &= \exp[\mu] \\ \text{Mode} &= \exp\left[\mu - \sigma^2\right]. \end{aligned} \tag{1.32}$$

Figure 1.6 shows the PDF plots of lognormal distributions for various values of μ and σ. As seen from these plots, the lognormal distribution is always skewed to the right: the PDF increases from zero to its mode and decreases thereafter. For a fixed μ, the PDFs skewness increases with σ; similarly, for a given σ, the degree of skewness increases when μ increases. Lognormal distributions arise in a variety of applications ranging from the insurance losses to the incubation period of an infectious disease.

1.1.7 Gamma Distribution

A gamma distribution is a two-parameter continuous probability distribution with PDF

$$f(t;\theta,\alpha) = \frac{\theta^{\alpha}}{\Gamma(\alpha)}\exp(-\theta t)\, t^{\alpha-1}, \tag{1.33}$$

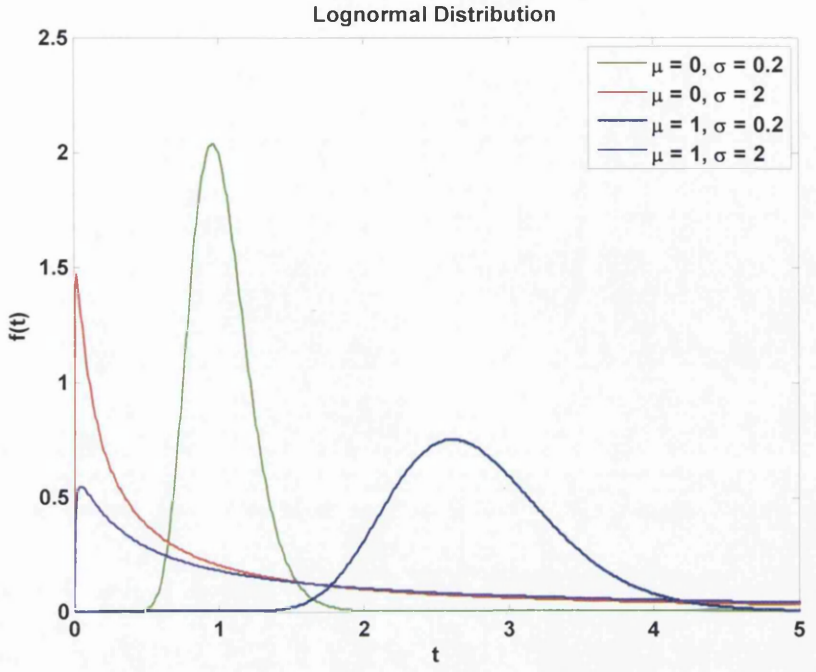


Figure 1.6: PDF plots of lognormal distributions for varying μ and σ.

for $t \geq 0$ where $\theta > 0$ is the rate parameter and $\alpha > 0$ is the shape parameter. The CDF is

$$F(t;\theta,\alpha) = \frac{\gamma(\alpha,\theta t)}{\Gamma(\alpha)} \tag{1.34}$$

where $\gamma(\alpha,\theta t)$ is the incomplete gamma function and the gamma function $\Gamma(\alpha)$ is defined in the next section. The Laplace transform of a gamma distribution is

$$L(s) = \frac{\theta^{\alpha}}{(s+\theta)^{\alpha}}. \tag{1.35}$$

The mean of a gamma distribution is given by

$$E[T] = \frac{\alpha}{\theta} \tag{1.36}$$

while the variance is

$$Var[T] = \frac{\alpha}{\theta^2} \tag{1.37}$$

and the mode is defined as

$$\text{Mode} = \frac{(\alpha-1)}{\theta} \text{for } \alpha \geq 1. \tag{1.38}$$

Note that there is no simple closed form for the median of a gamma distribution.

The shape parameter, α, determines the shape of the gamma distribution. From Figure 1.7, we note that when $\alpha > 1$, the distribution is bell-shaped, suggesting little rate het-

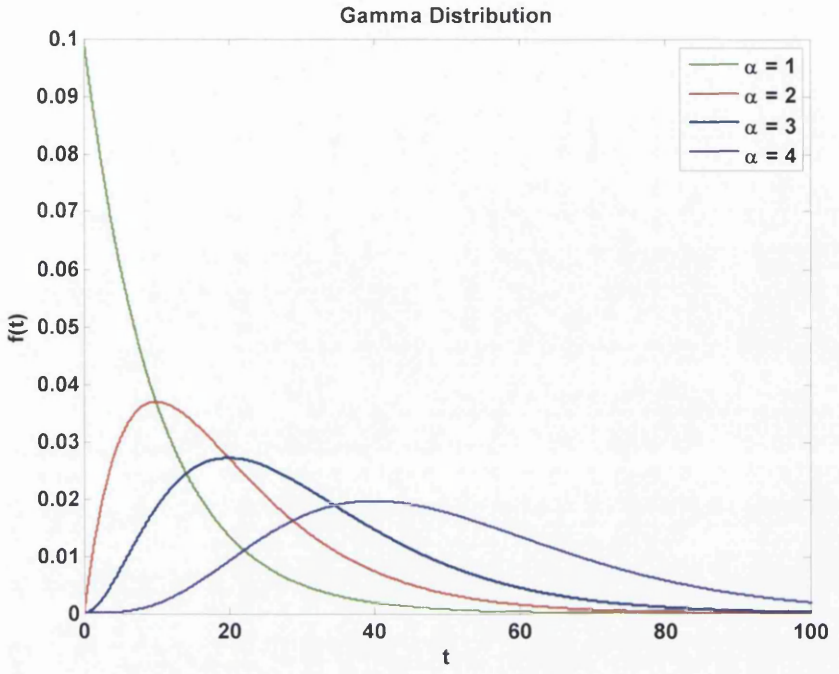


Figure 1.7: PDF plots of gamma distributions for varying α; $\theta = 0.1$.

erogeneity; the distribution is highly skewed when $\alpha < 1$ and is L-shaped, indicating high levels of rate variation. This flexibility makes the distribution suitable for accommodating different levels of rate variation in different data sets.

1.2 Useful Functions

In this section, we summarise some functions which are useful for developing the theory associated with mixture modelling.

1.2.1 Gamma Function

The Gamma function is defined by

$$\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t)\, dt \tag{1.39}$$

for $z > 0$. It is well known that

$$\Gamma(z+1) = z!$$

for integer values of z. We also know the recurrence relation

$$\Gamma(z+1) = z\Gamma(z).$$

The derivatives of the Gamma function can be expressed in terms of the Psi or Digamma Function, given by

$$\psi(z) = \frac{\partial[\ln\Gamma(z)]}{\partial z} = \frac{\Gamma'(z)}{\Gamma(z)}.$$

1.2.2 Pochhammer Symbol

The Pochhammer symbol

$$(x)_k = \frac{\Gamma(x+k)}{\Gamma(k)} = x(x+1)...(x+k-1) \tag{1.40}$$

for $k \geq 0$ is a notation for the rising factorial.

1.2.3 Hypergeometric Function

The generalised hypergeometric series is written as

$${}_pF_q(a_1,...,a_p;b_1,...,b_q;z) = \sum_{k=0}^{\infty} \frac{(a_1)_k ... (a_p)_k}{(b_1)_k ... (b_q)_k} \frac{z^k}{k!} \tag{1.41}$$

where $(a)_n$ is the Pochhammer symbol in (1.40). The classical standard hypergeometric series (when $p = 2$ and $q = 1$), given by

$${}_2F_1(a_1,a_2;b_1;z) = \frac{\Gamma(b_1)}{\Gamma(a_1)\Gamma(a_2)} \sum_{k=0}^{\infty} \frac{\Gamma(a_1+k)\Gamma(a_2+k)}{\Gamma(b_1+k)} \frac{z^k}{k!} \tag{1.42}$$

is useful for the computation of the variance of moment estimator for the discrete mixture of geometric distributions.

1.3 Kolmogorov-Smirnov Test

In several instances in this thesis, we wish to compare the fit to data of proposed distributions. The Kolmogorov-Smirnov test (hereafter abbreviated to KS test), a powerful test for goodness of fit, makes use of the CDF, $F(t)$ and the ECDF, $F_n(t)$ to tell us how well a distribution fits to a data set. In a KS plot, the ECDF is plotted with every jump being $\frac{1}{n}$, together with the CDF plot. The distance between the two functions are then found for every point. Before the jump, the ECDF is smaller than the one after the jump, so we find two distances KS^+ and KS^-. The distances between the ECDF before the jump and the CDF is KS^-, which is given by

$$KS^- = \max_i \left[F\left(t_{(i)}\right) - \frac{i-1}{n} \right]^+ \tag{1.43}$$

where $t_{(i)}$ are ordered from smallest to largest value. In general, with A^+ we mean

$$\begin{aligned} A^+ &= A \quad \text{if } A > 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

(1.43) means that if KS^- is negative, then we treat it as a zero. The distance between the ECDF after the jump and the CDF is KS^+, if KS^+ is negative, then we treat it as a zero. Therefore, the distance is

$$KS^+ = \max_i \left[\frac{i}{n} - F\left(t_{(i)}\right)\right]^+.$$

When we know both KS^- and KS^+, the Kolmogorov-Smirnov's distance is

$$KS = \max\left[KS^-, KS^+\right].$$

The closer is the estimated parameters to the true values, the shorter is the distance between the ECDF and the CDF.

An attractive feature of this test is that the distribution of the KS test statistic itself does not depend on the underlying CDF being tested. Another advantage is that it is an exact test (the Chi-squared goodness of fit depends on an adequate sample size for the approximations to be valid).

However, the KS test is only applicable to continuous distributions. It is less sensitive at the tails, compared to the centre of the distribution. The main drawback of this test is that the distribution must be fully specified. The critical region of the test is no longer valid if the parameters are estimated from the data.

1.4 Measures of Performance

We use three measures of performance to examine the performance of different estimators considered in this thesis; these are the square of bias in estimator, variance of estimator and the mean square error.

1.4.1 Square of Bias in Estimator

The bias of an estimator is given by

$$E\left[\hat{\theta}\right] - \theta$$

where θ is the true parameter and $E\left[\hat{\theta}\right]$ is the expected value. In our simulation process, we simulate 10000 samples and obtain 10000 estimates $\hat{\theta}$; we then find the bias using the average of these 10000 estimates, denoted as $\bar{\hat{\theta}}$ where

$$\bar{\hat{\theta}} = \sum_{i=1}^{10000} \frac{\hat{\theta}_i}{10000},$$

hence the bias is given by

$$\overline{\hat{\theta}} - \theta$$

We are concerned about the square of bias, which is

$$\left(\overline{\hat{\theta}} - \theta\right)^2.$$

1.4.2 Variance of Estimator

We also find the variance of the 10000 estimates $\hat{\theta}$ to tell how large is the deviation of the estimates from the true values. This is calculated from

$$Var_s\left[\hat{\theta}\right] = \sum_{i=1}^{10000} \frac{\left(\hat{\theta}_i - \overline{\hat{\theta}}\right)^2}{9999}.$$

1.4.3 Mean Square Error

The mean square error (MSE) is the sum of the square bias and the variance of estimator. It gives us an overview of the estimators incorporating both the bias and variance of estimators. It tells us how close are the estimates to the real value and how large is the variance of estimators at the same time. In our simulation results, we find the MSE of each method by

$$MSE_s\left[\hat{\theta}\right] = \left(\overline{\hat{\theta}} - \theta\right)^2 + Var_s\left[\hat{\theta}\right]$$

and use them for comparison means.

1.5 Finite Mixture Model

Over a century ago, Karl Pearson (1894) was among the first persons who fitted a mixture of two normal PDFs to a data set, that consisted of measurements on the ratio of forehead to body length of crabs sampled from the Bay of Naples. The data was provided by his colleague Weldon who speculated that the reason for the asymmetry in the histogram of the crab data was due to the existence of two new subspecies in the population. Ever since Pearson's classic paper was published, mixture models have attracted great attention and have wide applications in different fields, for instance, astronomy, biology, epidemiology, economics, engineering, marketing and medicine. The flexibility of finite mixture models makes them useful for modelling the heterogeneity in data, unknown distributional shapes and complex distributions. This is why we see the enormous expansion in recent decades of the literature about mixture modelling.

1.5.1 Basic Definition

Let $t_1, ..., t_{n_o}$ be the observed values of a random sample of size n_o. In fitting a finite mixture of m components to these data, it is assumed that the PDF of *the underlying random variable* t can be represented in the form of

$$f(t;\Theta) = \sum_{j=1}^{m} p_j f_j(t;\theta_j) \tag{1.44}$$

where $j = 1, ..., m$, $0 \leq p_j \leq 1$ and $\sum_{j=1}^{m} p_j = 1$. The $f_j(t;\theta_j)$'s are called the component densities of the mixture; whereas the p_j's are called the mixing proportions or weights. We let

$$\Theta = (\boldsymbol{\theta}, \boldsymbol{p})$$

where $\boldsymbol{p} = (p_1, ..., p_{m-1})$ is the vector of mixing proportion and $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$ denotes the vector of all unknown parameters of the component densities.

1.5.2 Modality

Finite mixture models are useful in representing the heterogeneities in an observed sample. The shape of density is very flexible in mixture models. One would expect a sample that arises from a mixture of distributions to have more than one mode. However, a mixture of distributions is not necessarily multimodal. In Figure 1.8, we see the PDF plots of mixtures of two normal distributions with different parameters. Note that when $\mu_1 = 10$, $\sigma_1 = 3$, $\mu_2 = 15$, $\sigma_2 = 3$ and $p = 0.6$, the mixture distribution has only one mode. In many cases, samples with mixtures of distributions are unimodal. For instance, the mode of a mixture of exponentials is zero (see Figure 1.9). As Bhattacharya (1967) had pointed out, we cannot rely on the modality of the histogram to tell whether a sample has arisen from a mixture distribution.

1.5.3 Identifiability

Identifiability addresses the theoretical question of whether it is possible to uniquely estimate a parameter from a sample, however large. A mixture distribution is identifiable if and only if for all

$$\{f(t_i;\Theta) : \Theta \in \Omega\},$$

where Ω is the specified parameter space,

$$f(t_i;\Theta) = f(t_i;\Theta^*), \tag{1.45}$$

implies that

$$\Theta = \Theta^*. \tag{1.46}$$

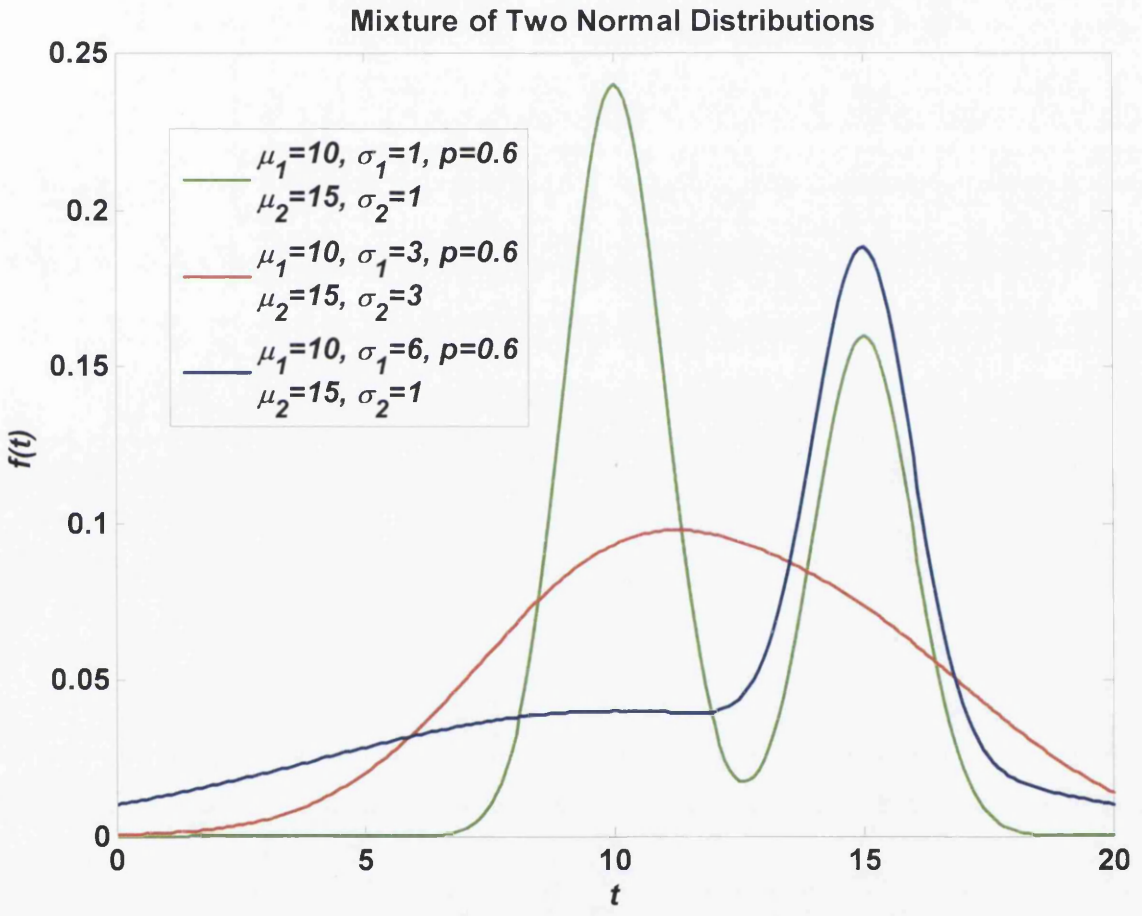


Figure 1.8: PDF plots of mixtures of two normal distributions: A normal mixture model can be unimodal.

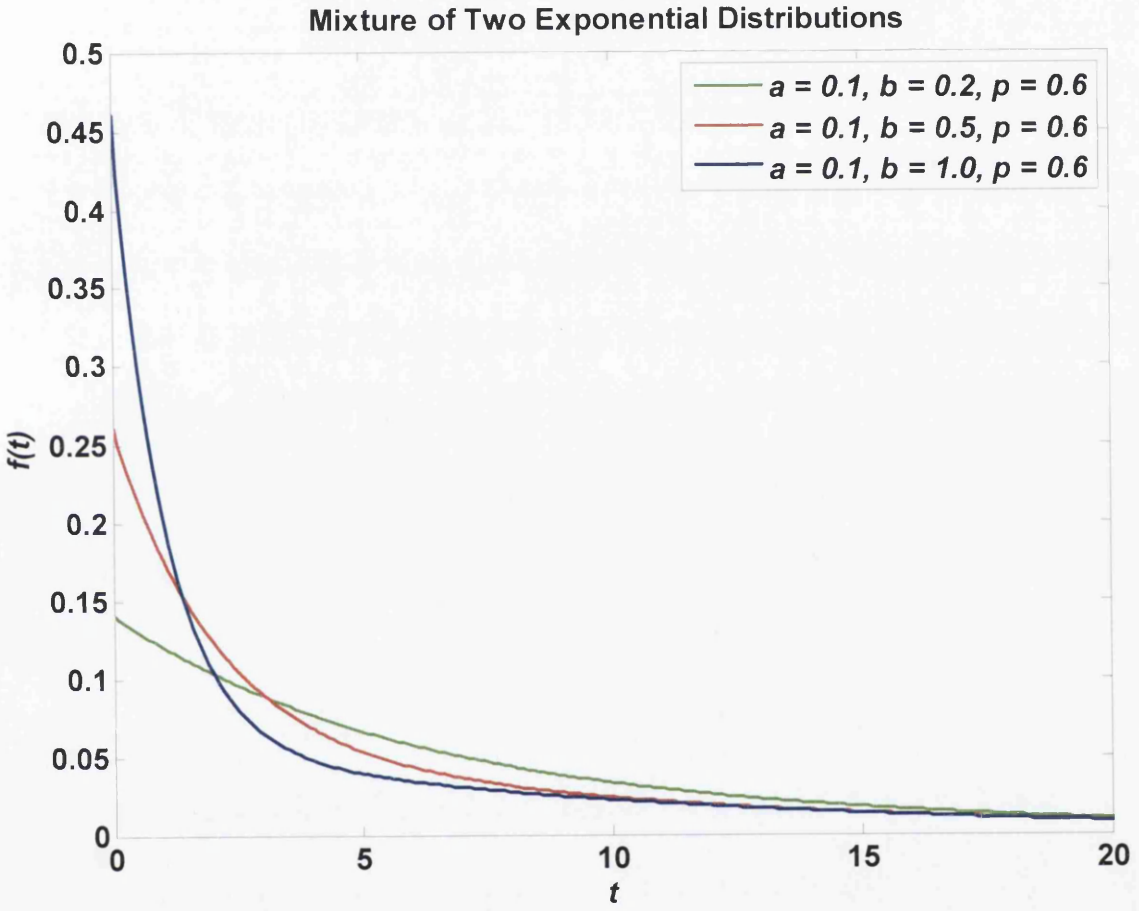


Figure 1.9: PDF plots of mixtures of two exponential distributions: An exponential mixture model is always unimodal.

Without identifiability, the estimation procedure for parameters Θ is meaningless. Suppose we have a mixture distribution with two component densities, say $f_1(t;\theta_1)$ and $f_2(t;\theta_2)$, that belong to the same parametric family. If we interchange the component labels 1 and 2 in Θ, (1.45) will still hold. In this case, however, Θ is not identifiable. If all the m component densities belong to the same parametric family, then $f(t_i;\Theta)$ is invariant under the $m!$ permutations of the component labels in Θ. This lack of identifiability is known as the label-switching problem in a Bayesian framework where posterior simulation is used to make inferences from the mixture model. For more information about the identifiability of a mixture distribution, readers are referred to Frühwirth-Schnatter (2006), Section 1.3.

1.5.4 Estimation Method

It is almost impossible to express the parameter estimates of mixture distributions in explicit forms. This is why, over the years, statisticians have paid tremendous efforts in simplifying the complications in mixture modelling. One can find a vast literature on estimation methodology for mixture distributions. A variety of estimation methods has been considered and compared, with the hope of obtaining the best candidate for estimating the parameters of mixture distributions with least errors.

We can imagine that the estimation problem of mixtures must have been much more difficult prior to the advent of high speed computers. Most of the estimation problems were solved using the method of moments (see Pearson (1894), Rider (1961), Cohen (1967)). The method of moments estimates the parameters by equating the sample moments with unobservable theoretical moments. For a mixed exponential distribution with m components, one needs to solve $2m-1$ moment equations for the estimates of $2m-1$ parameters. It has been proven that the efficiency of the method of moments can be greatly improved by using fractional moments (see Tallis & Light (1968)).

Graphical methods provide users with a quick but informal way to estimate mixtures of distributions. In the past, histograms, PDF plots, empirical CDF plots and so on have been used to obtain rough estimates of the parameters (see Bhattacharya (1967), Harris (1968), Fowlkes (1979)).

The maximum likelihood equations of mixture distributions are too complex to solve without the assistance of a computer. But since the advent of the Expectation-Maximisation Algorithm by Dempster *et al.* (1977), the maximum likelihood estimation for mixture models has become straightforward and hence is widely used by statisticians in solving mixture problems.

Other methods such as the minimum distance estimator have been considered, for instance, the Kullback-Leibler distance (Kullback & Leibler (1951)) between the CDF of the mixture distribution and the empirical CDF that places mass one at each data point, or other distances such as the Cramér-von Mises distance (Woodward *et al.* (1984) and Lindsay (1994)), the Kolmogorov distance (Deely & Kruse (1968)), the Hellinger distance (Karlis & Xekalaki (1998))

Less attention had been paid to the Bayesian approaches for estimation of mixture distributions until the publication of the key paper by Gelfand & Smith (1990), that showed that Gibbs Sampling has great potential in solving the mixture problems. Since then, the Bayesian approach has become another favourite choice for mixture modelling.

The main part of this thesis is about the use of the generalised method of moments for estimating mixtures of exponential distributions and the discrete analogue, mixtures of geometric distributions. Simulation experiments are carried out so that we can compare the performance of these methods to the MLE.

1.5.5 Asymptotic Covariance Matrix of Generalised Moment Estimator

In this thesis, we consider a number of methods which use other forms of moments to estimate the three parameters of a two-component exponential mixture and its discrete counterpart, geometric mixture. However, for mixture distributions, it is impossible to explicitly compute the asymptotic covariance matrix of moment estimator. In order to appraise the efficiency of a generalised moment estimator, we need to have a method to find at least an estimate of the variance of the estimator. Jalali (2006) constructed a general method to calculate the variance of the generalised moment estimator approximately. This method can be applied to the methods using generalised moments which will be discussed in the future chapters.

A mixture of m exponential distributions has a $m-1$ vector $\boldsymbol{p} = (p_1, ..., p_{m-1})$ of mixing weights and a m vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$ of distinct parameters. These estimators are functions of an α vector ($\alpha \geq 2m-1$) $\hat{\boldsymbol{\mu}}$ whose components are estimators of different generalised moments.

For simplicity, we first assume that $\alpha = 2m-1$, and define the $\alpha \times \alpha$ matrix $\boldsymbol{V}[\hat{\boldsymbol{\mu}}]$ as the covariance matrix of the generalised moments.

$$\boldsymbol{V}[\hat{\boldsymbol{\mu}}] = \{V_{ij}\}$$

where

$$V_{ij} = Cov\left[\hat{\mu}_i, \hat{\mu}_j\right].$$

Obviously, this matrix is a function of $\boldsymbol{p}$ and $\boldsymbol{\theta}$. Next, we let $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ be a $(2m-1)\times(2m-1)$ covariance matrix of $\hat{\boldsymbol{\Theta}} = \left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}\right)^T$. The component j of $\hat{\boldsymbol{\Theta}}$ is denoted by $\hat{\boldsymbol{\Theta}}_j$, where

$$\hat{\boldsymbol{\Theta}}_j = \boldsymbol{f}_j\left(\hat{\boldsymbol{\mu}}\right).$$

If we Taylor expand $\hat{\boldsymbol{\Theta}}$ about the mean of $\hat{\boldsymbol{\mu}}$ and consider its first two terms, then

$$\boldsymbol{Y} = \boldsymbol{f}\left(E\left[\hat{\boldsymbol{\mu}}\right]\right) + \Delta\left(\hat{\boldsymbol{\mu}} - E\left[\hat{\boldsymbol{\mu}}\right]\right),$$

where $\triangle$ is the Jacobian matrix with entries $\triangle_{ij} = \dfrac{\partial \boldsymbol{f}_i}{\partial \hat{\boldsymbol{\mu}}_j}$. It then follows that

$$\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right] = \triangle \boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right] \triangle^T.$$

This is just the well-known method of approximating the variance of functions of a random vector. The slight snag here is the fact that functions f_i's are mostly implicit.

To find the Jacobian matrix $\triangle$ we scrutinise how we obtained our $\hat{\boldsymbol{\Theta}}_j$'s. We first found an expression in terms of our parameters $\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}\right)^T = \hat{\boldsymbol{\Psi}}$ for the vector $\hat{\boldsymbol{\mu}}$. We therefore had explicit expression of the form

$$\hat{\boldsymbol{\mu}} = \boldsymbol{g}\left(\hat{\boldsymbol{\Psi}}\right).$$

Then we inverted this as

$$\boldsymbol{\Psi} = \boldsymbol{g}^{-1}\left(\hat{\boldsymbol{\mu}}\right)$$

and wrote

$$\hat{\boldsymbol{\Theta}} = \boldsymbol{g}^{-1}\left(\hat{\boldsymbol{\mu}}\right).$$

This means that

$$\boldsymbol{f} = \boldsymbol{g}^{-1},$$

i.e. f is the functional inverse of g. Now, if we denote the Jacobian of g by $\triangle$, then we know from Mathematical Analysis that

$$\triangle = \boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1},$$

where the ij^{th} entry of the Jacobian $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$ are $d_{ij} = \dfrac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{\Theta}_j}$. Hence,

$$\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right] \approx \boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1} \boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right] \left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1}\right)^T. \tag{1.47}$$

To summarise: in order to find the approximation for the covariance matrix $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ of the estimators we should take the following five steps.

1. Find the covariance matrix $\boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right]$ of our generalised moments.
2. Write explicit formulae $\mu_k = g_k\left(x\right)$ expressing the generalised moments in terms of the parameters $\boldsymbol{\Psi}$.
3. Find the Jacobian matrix of functions, $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$.
4. Invert the matrix $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$ and evaluate it at $\hat{\boldsymbol{\Psi}}_i = \hat{\boldsymbol{\Theta}}_i$.
5. Now use the covariance formula in (1.47).

1.6 Hidden Markov Models

A hidden Markov model (HMM) is a stochastic process generated by a Markov chain whose state sequence cannot be observed directly, only through a sequence of observations. The inter-event time distribution of an aggregated Markov process is a linear combination of exponential distributions with different rates. This is the main reason for our interest in the mixture of exponential (and of geometric) distributions (see Section 1.7).

One of the examples of HMM that is considered in this thesis is a model for spread of infectious agents between hosts suggested by Gravenor (2003). The particular disease system that we investigate here is that of prion diseases, specifically scrapie.

1.6.1 Scrapie Disease

Scrapie is the canonical member of the transmissible spongiform encephalopathy (TSE) group of diseases. It is a disease of sheep known in the UK for over 250 years. It is an infectious, and invariably fatal disease that is transmitted naturally between sheep. Scrapie causes itching in infected animals, leading to compulsive rubbing and loss of hair. Symptoms are followed by severe neurodegeneration and death. A key characteristic is the abnormally long incubation period, which can be several years (Detwiler (1992)).

The Link to Kuru

Scrapie was eventually linked to a series of rare and unusual diseases of humans, now also known to be TSEs. Creutzfeld-Jacob Disease (CJD) was a rare sporadic or genetic degenerative disorder first described in 1920. A disease with similar symptoms, known as kuru, was then identified in the Fore tribespeople of Papua New Guinea in the 1950s. Both the diseases are fatal, incurable and largely untreatable. The crucial link between the veterinary disease scrapie and the new human disease kuru was made by Hadlow (1959). Epidemiological work in Papua New Guinea, had led to suspicions that kuru was not sporadic or genetic (like CJD) but was caused by an infectious agent and spread by cannibalistic tribal rituals whereby the Fore people consumed the remains (crucially including the brain) of relatives, including those who had died from kuru. After Hadlow's suggestion, Gajdusek *et al.* (1966) demonstrated experimentally that kuru, like scrapie, was indeed transmissible. Cannibalistic practices became very rare amongst the Fore people after the 1950s, and kuru began to die out, yet due to the long incubation period, new cases are still occasionally encountered today.

Mad Cow Disease and vCJD in Humans

After the curiosity of the kuru epidemic, TSEs remained rare and unusual diseases that were not intensively studied. Similar diseases were noted in mink and deer (chronic wasting disease, or CWD). The infectious agent or parasite remained mysterious and was usually

classified as an "unconventional" or "slow" virus. New interest was sparked by the sudden and dramatic appearance of a new disease of cattle in the UK in 1986: bovine spongiform encephalopathy (BSE) or "mad cow disease" (Wilesmith *et al.* (1988)). Studies quickly showed the similarities in neurodegeneration and symptoms to other TSEs, and epidemiological investigations revealed the likely route of transmission: like kuru, the "cannibalistic" practices of feeding cows with protein supplemented with recycled meat and bone meal obtained from rendered carcasses from the cattle industry. The sterilisation process (heating at pressure) of rendering carcasses was designed to destroy bacteria and viruses but was clearly insufficient to inactivate the TSE agent. Due to the long incubation period, infected but apparently healthy cows sent for slaughter were rendered for use in feed. In this manner one infected animal could infect many other cattle exposed to the feed. Several cycles of amplification of infection had already taken place by 1986, and many thousands of cattle were already incubating the disease. Despite prompt bans of the feed practice (and later reinforcements of the bans) over 170,000 clinical cases were eventually detected, with several millions of UK cattle likely to have become infected (Anderson *et al.* (1996)). Exports (before export bans) lead to cases in almost all European countries, and outside Europe including USA and Canada (European Food Standards Agency: http://www.efsa.europa.eu).

The BSE crisis was initially a veterinary and economic challenge. But from the early days there was the concern that the new disease might be transmitted to humans. The long held assumption that scrapie was harmless to humans reduced fears. However, in 1994 a new or "variant" form of CJD appeared. Confirmation that BSE had transmitted from cattle to humans was established, by scrapie experts, experimentally in 1996 (Bruce *et al.* (1997)) and fears began of a potentially large epidemic of vCJD in the UK, due to the large number of infected cattle that had been consumed. Cases of vCJD began to rise, but worst-case scenarios luckily have not arisen and it appears that humans are much less susceptible to BSE than feared (see "species barrier" below). To June 1st 2009, 164 deaths due to vCJD have been recorded (UK CJD Surveillance Unit, Edinburgh, http://www.cjd.ed.ac.uk/figures.htm), with only 1 death in 2008 and 2009. The outbreak in humans appears to have peaked in 2000 (28 deaths) and predictions (obtained from mapping the BSE outbreak in cattle to human exposure) suggest a maximum final outbreak of several hundred cases.

The Prion Theory

The importance of the BSE-CJD link sparked huge research interest in TSEs, with a focus on identifying and characterising the infectious agent. Prusiner (1982) had proposed the "prion theory" (Prusiner (1982)) for all TSEs. The idea was revolutionary, that unlike other transmissible infectious disease, the agent was not a virus, bacteria or parasite, but a simple "proteinaceous infectious agent". Hence there was no nucleic acid (DNA or RNA) to encode the replication of the agent. Instead Prusiner suggested that a naturally occurring cellular protein, termed PrP existed in two states. In the normal, healthy state PrPc does no damage (indeed it must fulfil a useful metabolic function). However, the same

protein is able to change shape into a new conformation PrPSc which, if it accumulated (for example in the brain) causes pathological damage leading to symptoms typical of TSEs. Crucially, PrPSc appears to catalyse the conversion of PrPc into PrPSc. Therefore if PrPSc is introduced into a host (via an infection process such as cannibalism, blood transfusion or contamination) the abundance of the scrapie form of PrP will increase gradually leading to disease and eventual death. Any transmission of material containing PrPSc from this host can potentially continue the chain of infection. There is now compelling evidence to support this hypothesis that the TSE infectious agent is composed exclusively by a misfolded version of the prion protein that replicates in the body in the absence of nucleic acids by inducing the misfolding of the normal cellular prion protein (although the process has not been fully demonstrated, see below).

The Species Barrier and Sub-Clinical Infection

For prion disease, the transmission between different species is usually limited by the "species barrier" (Pattison (1965)). For instance, it takes a far greater dose of cow-derived BSE inoculum to achieve equal infection rates in mice compared to cow transmission. Such a species barrier increases the incubation period and reduces the percentage of other species animals that succumb to disease (Hill (2000)). The recipient which shows no symptoms of disease is called "resistant" species. Recently, researches have raised the possibility of "sub-clinical" infection of scrapie among animals. Experiments of scrapie disease in mice have shown that sub-clinically infected mice will not experience symptoms of scrapie, but are infectious on further transmission. The author is interested in estimating the prevalence of sub-clinical infection in the experimental group of mice using a special Markov model. The results are shown in Chapter 7.

1.6.2 Analysis of Serial Passage PrP^{Sc} Experimental Data

The author's work is motivated by a project which deals with de novo generation of infectious mammalian prions from recombinant prion protein. Despite the award of the Nobel Prize to Prusiner for his prion theory in 1997 (following on from Carelton Gajdusek's Nobel Prize in 1976 for confirming the infectious nature of kuru in humans) there still remains some controversy over the exact nature of the infectious agent: whether it is purely proteinaceous in origin, or whether there remains some viral (nucleic acid) involvement. Recent work to "complete" the prion theory has focused on whether PrPc (the normal form of the protein) can be induced under experimental conditions in the laboratory to behave like PrPSc (the infectious form). There have been some successes (see Barria *et al.* (2009)), however the demonstration requires both the presence of scrapie disease after inoculation and its further transmission to other hosts. Because the generation of de novo prions appears to be a rare event, experimental system needs to be very sensitive to allow comparison to a control system. A further complication is that many infectious agents, in particular prion

diseases that are characterised by extremely long incubation periods, may cause "sub-clinical infection", which only manifests itself on subsequent passage when transmitted to a new host who then displays symptoms. These questions motivated the need for an appropriate statistical model for the "serial passage" of prion disease to help test the prion hypothesis in the most important experimental system: scrapie in rodents.

In an experimental system of scrapie disease in mice, the waiting time for the host to exhibit signs of scrapie can be modelled by a special kind of Markov process. The aim of the project is to "track" the sub-clinically infected animals, giving an estimate of the overall prevalence of infection. The experimental data for the investigations were kindly supplied by Professor Adriano Aguzzi and his laboratory at the Institute of Neuropathology, University of Zurich.

Figure 1.10 presents the structure of the typical serial passage experiment. In the experimental group H10 recPrP$^{\beta}$, ten mice are exposed at first passage (indicated by green arrows) to a recombinant prion protein. None of the primary passage mice develop scrapie disease. Six of the primary passage mice are chosen at random for second passage (indicated by orange arrow) into either three or four mice. Scrapie disease is detected in one of the contacts (indicated by red circle), and successfully transmitted to a further four mice. The predecessor can now be revealed as sub-clinically infected (indicated by purple circle). The implication is that recPrP causes a sub-clinical infection on first passage, which can manifest as disease on subsequent passage. Of course, this raises the question of just how many sub-clinically infected animals are present at first passage?

If an exposed mouse shows no symptoms of disease, this does not necessarily mean that it is not infected by the prion protein. This is because a sub-clinically infected mouse also does not show symptoms of scrapie disease. Therefore one can hardly distinguish a healthy mouse and a sub-clinically infected mouse. If the experiments are carried on for a large number of serial passages, we would know the prevalence of scrapie at the first passage. However, the experiment is extremely time consuming (sometimes taking several years to complete) and costly and hence a reliable method to predict the prevalence scrapie at first passage by using a statistical model is required.

Generally, we propose a Markov process where the mice exist in one of three states, which are "Diseased" (definitely infected, showing clinical symptoms) (D), "Uninfected and Healthy" (H), and "Sub-clinically Infected" (S) , and transition between all states is theoretically possible, although we will impose some biologically plausible constraints. The states refer to the condition of the mouse at the time it is used to initiate further passage. In reality, diseased hosts always cause disease on serial passage and uninfected hosts cannot cause disease. Therefore, the transition probabilities matrix is:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ p_{sd} & p_{sh} & 1 - p_{sd} - p_{sh} \end{pmatrix}$$

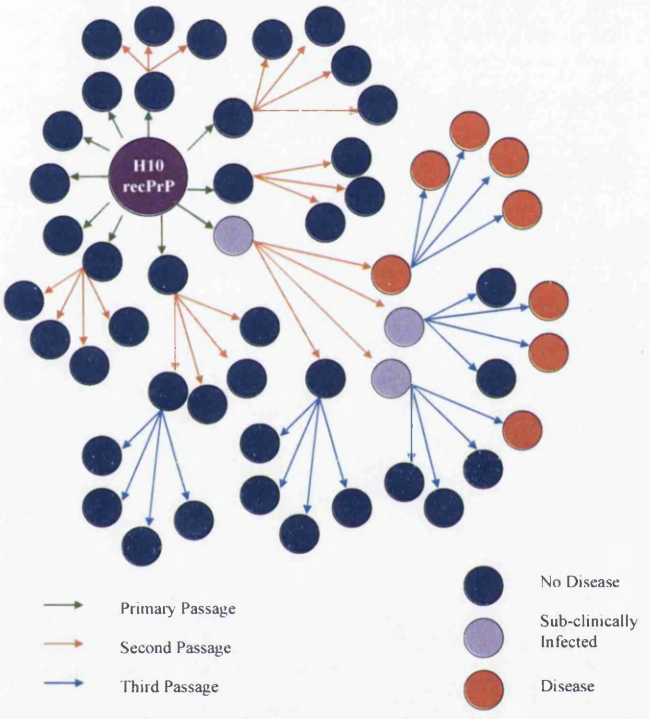


Figure 1.10: H10 recPrP$^{\beta}$ experimental data and design of a typical "serial passage" experiment. Data provided by Professor A. Aguzzi, Institute of Neuropathology, University of Zurich.

where p_{sd} is the probability of a S to D transition and p_{sh} is the probability of a S to H transition.

One aim of the model is to provide evidence for a highly significant increase in the probability of prion disease arising during serial transfer in a series of mice given recPrP on primary passage (the H10 recPrP$^{\beta}$ group) compared to the set of controls. However our general point of interest is to "track" the sub-clinically infected animals and give greater biological insight.

As suggested by Jalali (2005c), if we clump the "Healthy" state (H) and "Sub-clinically Infected" (S) state into one level, which is named as "No-Disease" (E), the waiting time in the "No-Disease" state has a mixture of two geometric distributions. In his note, Jalali has completely solved the problem of two hidden state discrete Markov cases. He has also given instructions for estimation of parameters and reconstruction of Markov transition matrix. In later sections, the proof that mixture of geometric distributions is analogous to mixture of exponential distributions will be given. This means that all the methods (as discussed before) used for estimating the parameters of a mixture of two exponential distributions can be applied in a similar way to the distribution of the waiting time of the "No Disease" state (E). Having found the parameters of the distribution of the waiting time (i.e. a mixture of two geometric distributions), we can then estimate the transition parameters of the prion model. This is the topic of Chapter 7.

Lastly, we note that the model has applications for the study of sub-clinical infection in the epidemiological setting. For several important diseases such as tuberculosis, sub-clinical infection cannot be detected but can be transmitted. For the epidemiology model, some structure is used, but the constraints of the transition matrix are relaxed. We consider the properties of the epidemiology model in Chapter 8.

1.7 Clump Model

In practice, we might not be able to distinguish between states of the Markov process. For instance, in the prion model discussed above, the two states "Healthy"(H) and "Sub-clinically Infected"(S) are indistinguishable. A host shows no symptoms of disease when it is sub-clinically infected. We can only reveal that a host is sub-clinically infected when its successor shows disease. The state of a non-diseased host remains mysterious unless its successor shows symptoms of scrapie disease. In such a situation, we could suggest grouping these two states together into one level representing "No Disease" (E).

When we clump the indistinguishable states into one state, the distribution of waiting time is no longer a pure exponential distribution but a mixture of exponential distributions. Note that this is for a continuous process. For the prion model, the waiting time is discrete, so it has a mixture of geometric distributions.

1.7.1 The Problem

In his note, Jalali (2005c) showed how his work on mixtures of Polya-Lagguer Finite Class (Jalali (2002)) can be adapted to the discrete case with minimal effort. The following results were proposed in his paper. Suppose we have an $(m+1)$-state Markov chain (time homogenous and discrete), in which we have isolated one state and lumped the rest of the states into one "observable" entity, which we call a "level". The transition matrix of this Markov chain is as follows:

$$\begin{bmatrix} \alpha & \boldsymbol{u} \\ \boldsymbol{v} & \boldsymbol{P} \end{bmatrix}, \tag{1.48}$$

where α is a scalar probability not equal to 1, $\boldsymbol{u}$ is a row m-vector of probabilities, $\boldsymbol{v}$ is a column m-vector of probabilities, and $\boldsymbol{P}$ is an m by m matrix.

Let N be a discrete random variable which represents the "lifetime" in our level from the time the level is entered into from state 1. What we do is investigating the properties of the distribution of this random variable with the ultimate aim of collecting enough information about it to enable us to estimate its parameters.

The probability vector of entry into the level is $\frac{\boldsymbol{u}}{1-\alpha}$. N takes every positive integer value, and the probability mass at n equals

$$f(n) = \frac{\boldsymbol{u}}{1-\alpha}\boldsymbol{P}^{n-1}\boldsymbol{v}. \tag{1.49}$$

It can be shown that the generating function of N has the following simple expression:

$$G(s) = \frac{\boldsymbol{u}s}{1-\alpha} \times \frac{\boldsymbol{I}-\boldsymbol{P}}{\boldsymbol{I}-s\boldsymbol{P}} \times \mathbf{1}_m. \tag{1.50}$$

(In (1.50) the matrix fraction is not ambiguous because the numerator and the denominator commute.)

The Analogous Problem:

Instead of a discrete chain, suppose we now have a continuous time Markov process. We similarly partition its generating matrix as follows:

$$\begin{bmatrix} -\beta & \boldsymbol{u}_c \\ \boldsymbol{v}_c & \boldsymbol{Q} \end{bmatrix}, \tag{1.51}$$

where β is a positive scalar, $\boldsymbol{u}_c$ is a non-negative row m-vector, $\boldsymbol{v}_c$ is a non-negative m-vector, and $\boldsymbol{Q}$ is an $m \times m$ matrix.

We assume, without any loss of generality, that the modulus of each element of this generator does not exceed 1. The fact that we do not lose generality is because we can always choose our unit of time small enough to guarantee this requirement. Now let T denote the continuous random variable which represents the lifetime in the level from the

time this level is entered into from state 1, to the time of its exit. We know that the survival function of T equals

$$S(t) = \frac{\boldsymbol{u}_c}{\beta} (\exp \boldsymbol{Q}t) \mathbf{1}_m. \tag{1.52}$$

The PDF of T is therefore

$$f(t) = \frac{\boldsymbol{u}_c}{\beta} (-\boldsymbol{Q}) (\exp \boldsymbol{Q}t) \mathbf{1}_m, \tag{1.53}$$

so that the Laplace transform of (1.53) is given by

$$L(s) = \frac{\boldsymbol{u}_c}{\beta} \times \frac{-\boldsymbol{Q}}{s\boldsymbol{I} - \boldsymbol{Q}} \times \mathbf{1}_m. \tag{1.54}$$

The Analogy:

We next replace in (1.50) s by $\dfrac{1}{1+s}$, and $\boldsymbol{I} - \boldsymbol{P}$ by $-\boldsymbol{Q}$. This gives us

$$\begin{aligned} & \frac{\boldsymbol{u}}{1-\alpha} \times \frac{1}{1+s} \times \frac{-\boldsymbol{Q}}{\boldsymbol{I} - \dfrac{(\boldsymbol{I}+\boldsymbol{Q})}{1+s}} \times \mathbf{1}_m \\ = & \frac{\boldsymbol{u}}{1-\alpha} \times \frac{-\boldsymbol{Q}}{s\boldsymbol{I} - \boldsymbol{Q}} \times \mathbf{1}_m. \end{aligned}$$

With these substitutions (transformation), the generating function of N become exactly the same as the Laplace transform of T. So the study of almost all properties of N is analogous to the study of the corresponding properties of T. In other words, any results of the study on T can be rephrased for N. Note that with our assumption about the size of transition intensities of our Markov process, $\boldsymbol{I} + \boldsymbol{Q}$ is legitimate $\boldsymbol{P}$, and conversely $\boldsymbol{P} - \boldsymbol{I}$ a legitimate $\boldsymbol{Q}$ satisfying our assumption. Similarly, $\boldsymbol{u}_c$ has the same properties as $\boldsymbol{u}$, $\boldsymbol{v}_c$ as $\boldsymbol{v}$ and $1 - \alpha$ as β, and vice versa. We can summarise the foregoing facts in the following theorem.

Theorem 1 *(Jalali (2005c)) Let $T(\beta, \boldsymbol{u}_c, \boldsymbol{v}_c, \boldsymbol{Q})$ and $N(\alpha, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{P})$ be random variables defined as above. Let further $\beta = 1 - \alpha$, $\boldsymbol{u}_c = \boldsymbol{u}$, $\boldsymbol{v}_c = \boldsymbol{v}$, and $-\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{P}$. Then we have*

$$E[\exp(-sT)] = E\left[(1+s)^{-N}\right],$$

and

$$E[\theta^N] = E\left[\exp\left(-\frac{1-\theta}{\theta}T\right)\right].$$

By the k^{th} falling factorial moment (FFM) of a random variable X, we mean

$$E[X(X-1)...(X-k+1)] \tag{1.55}$$

and by the k^{th} rising factorial moment (RFM) of X we mean

$$E\left[X\left(X+1\right)...\left(X+k-1\right)\right]. \tag{1.56}$$

Let us assume the parameter identities expressed in the theorem all hold, so when we speak of T and N we know what parameters they are based on and we know the relationships between the two sets of parameters. So we safely dispense with specifying the parameters of these random variables explicitly. We denote by μ_k and μ_k^* the k^{th} (ordinary) moments of T and N respectively; by ϕ_k and ϕ_k^* the k^{th} FFMs of T and N; and ρ_k and ρ_k^* the k^{th} RFMs of T and N. We next prove this consequence of Theorem 1.

Theorem 2 *(Jalali (2005c)) For every k, we have*

1. $\mu_k = \rho_k^*$
2. $\mu_k = \sum_{i=0}^{k} S_{k,i}\mu_i^*$, *where $S_{k,i}$ are Stirling numbers of the second kind;*
3. $\mu_k^* = \sum_{i=0}^{k} s_{k,i}\mu_i$, *where $s_{k,i}$ are Stirling numbers of the first kind;*
4. $\mu_k^* = \phi_k$

Proof.

1 This is proved by differentiating both sides of the first equation of Theorem 1*:*

$$\begin{aligned}
\mu_k &= (-1)^k \left[\frac{d^k}{ds^k} E\left[\exp\left(-sT\right)\right]\right]_{s=0} \\
&= (-1)^k \left[\frac{d^k}{ds^k} E\left[\left(1+s\right)^{-N}\right]\right]_{s=0} \\
&= \left[E\left[N\left(N+1\right)...\left(N+k-1\right)\left(1+s\right)^{-N-k}\right]\right]_{s=0} \\
&= \rho_k^*.
\end{aligned}$$

2 It is well known that

$$N\left(N+1\right)...\left(N+k-1\right) = \sum_{i=0}^{k} S_{k,i}N^i.$$

The desired identity follows from 1 *by finding the expectaion of both sides.*

3 This follows from 2 *by Stirling conversion formula.*

4 It is well known that

$$T\left(T-1\right)...\left(T-k+1\right) = \sum_{i=0}^{k} s_{k,i}T^i.$$

So by finding the expectation of both sides we have

$$\phi_k = \sum_{i=0}^{k} s_{k,i}\mu_i$$

The RHS, by 3, *is just* μ_k^*.

■

Now as result of Theorem 2, we can use any method of moments developed for estimating parameters of T, to estimate the parameters of N. The only difference is that everywhere we need an estimate from data of an ordinary moment for estimation of T, we provide the estimation for RFM from data for N.

Continuous Case

When $m = 1$, (1.54) is written as

$$L(s) = \frac{\theta_e}{s + \theta_e}, \tag{1.57}$$

where $\theta_e = -\boldsymbol{Q}$, which is a positive number and by our assumption $\theta_e \leq 1$. (1.57) is the Laplace transform of an exponential distribution.

When $m = 2$, the distribution of T is a linear combination (not necessarily a mixture) of two exponential distributions. The parameters of the exponential distribution are the eigenvalues of matrix $-\boldsymbol{Q}$, which are positive and real. The PDF has the form

$$f(t; a, b, p) = pa\exp(-at) + (1-p)\, b\exp(-bt), \tag{1.58}$$

where $t \geq 0$, $a < b$ and p is allowed to be greater than 1. If p is less than 1, then we have an essentially time reversible case, and the PDF is a mixture of exponentials. In any case, the mixing weight should satisfy the inequality $p \leq \frac{b}{b-a}$. In the marginal case where $a = b$, the PDF is a mixture of an exponential and a gamma distribution with shape parameter 2, and the same rate parameter as the exponential distribution. The estimation problem of a mixture of two exponential distributions is studied extensively in Chapter 3.

Discrete Case

Similarly, when $m = 1$ and let $\boldsymbol{P} = \theta_g$, (1.50) becomes

$$G(s) = \frac{(1-\theta_g)\, s}{1 - s\theta_g} \tag{1.59}$$

where $0 \leq \theta_g \leq 1$. (1.59) is the generating function of a geometric distribution. The connecting identity between (1.57) and (1.59) is $\theta_e = 1 - \theta_g$.

When $m = 2$, the generating function of N is by analogy

$$G(s) = p\frac{as}{1-(1-a)s} + (1-p)\frac{bs}{1-(1-b)s}. \tag{1.60}$$

Therefore, N has a mixture of two geometric distributions with PMF

$$f(n;a,b,p) = pa(1-a)^{n-1} + (1-p)b(1-b)^{n-1}, \tag{1.61}$$

where $n = 1,2,...$, $a < b$ and p is not necessarily less than 1. The parameter $\bar{a} = 1-a$ is the largest eigenvalue of $\boldsymbol{P}$ which is positive and not exceeding 1. $\bar{b} = 1-b$ is the other eigenvalue of $\boldsymbol{P}$ which is real and not greater than $\bar{a}$ in an absolute way, but may be negative. This eigenvalue is negative if and only if the determinant of $\boldsymbol{P}$ is negative. The fact that this eigenvalue can be negative adds a great deal of subtlety to the discrete case. First we note that (1.61) can be an actual PMF if it remains everywhere non-negative. As long as the mixing weight satisfies the following conditions:

$$p \leq \frac{1-\bar{b}}{\bar{b}-\bar{a}},$$

if $\bar{b} \geq 0$ and

$$\frac{-\bar{b}(1-\bar{b})}{(\bar{a}-\bar{b})} \leq p \leq \frac{1-\bar{b}}{\bar{b}-\bar{a}}$$

if $\bar{b} < 0$. So, given any PMF as in (1.61), we can construct matrices $\boldsymbol{P}$ whose associated N have a PMF equal to $f(n;a,b,p)$ in (1.61). When $\bar{b} \geq 0$, the construction will be exactly analogous to the continuous case. When $\bar{b} < 0$, we need a different type of construction. One such universal construction is the following:

$$\boldsymbol{P} = \begin{bmatrix} 0 & \dfrac{-\bar{a}\bar{b}}{1-\bar{a}-\bar{b}} \\ 1-\bar{a}-\bar{b} & \bar{a}+\bar{b} \end{bmatrix},$$

$$\boldsymbol{v} = \begin{bmatrix} \dfrac{(1-\bar{a})(1-\bar{b})}{1-\bar{a}-\bar{b}} \\ 0 \end{bmatrix},$$

and

$$\frac{\boldsymbol{u}}{1-a} = \begin{bmatrix} \dfrac{(1-\bar{a}-\bar{b})(1-\bar{b}-p(\bar{a}-\bar{b}))}{(1-\bar{a})(1-\bar{b})} & \dfrac{p(\bar{a}-\bar{b})(1-\bar{a}-\bar{b})+\bar{b}(1-\bar{b})}{(1-\bar{a})(1-\bar{b})} \end{bmatrix}.$$

The positive mixture of two geometric distributions is studied and presented in Chapter 4; whereas in Chapter 5, the linear combination of two geometric distributions is discussed.

1.7.2 Data Simulation

In Chapters 3 and 4, we will study a number of moment-based methods for estimating the parameters of an exponential mixture and a geometric mixture. Since one of our main purposes is to compare the performances of these estimators, we will need to simulate data from such distributions. We use Matlab for most computational tasks related to the simulation and analysis of data. In general, we can generate data t_i from a distribution using inversion methods, which are based on the observation that continuous CDFs range uniformly over the interval $(0, 1)$. If u_i are uniform random numbers on $(0, 1)$, and F is the CDF of a distribution, then

$$t_i = F^{-1}(u_i).$$

Thus, to simulate a set of data from an exponential distribution with a specified parameter θ, we use (1.2) and compute

$$t_i = -\frac{1}{\theta}\ln(1 - u_i).$$

For the discrete geometric, we can also use inversion methods to generate a set of data n_i using (1.8). We generate a uniform random number u_i on $(0, 1)$ and then set

$$N = n_i$$

if

$$F(n_i - 1) < u_i < F(n_i).$$

To generate a data set following a mixture distribution, we also need Bernoulli random numbers. A Bernoulli random number n_i takes value 1 with success probability p and value 0 with failure probability q $(= 1 - p)$. If we generate one uniform random number on the interval $(0, 1)$, n_i is 1 if it is less than p.

We use the following random number generators in Matlab to generate data sets:

$\boldsymbol{T}$ = exprnd(μ, m, n) generates random numbers from the exponential distribution with mean parameter μ $\left(= \theta^{-1}\right)$, where scalars m and n are the row and column dimensions of $\boldsymbol{T}$.

$\boldsymbol{N}$ = geornd(θ, m, n) generates geometric random numbers with probability parameter θ, where scalars m and n are the row and column dimensions of $\boldsymbol{N}$.

$\boldsymbol{B}$ = binornd(N, p, m, n) generates an m-by-n matrix containing random numbers from the binomial distribution with parameters N and p.

With these built-in functions, we generate, say, a mixture of two exponential distributions with specified parameters a, b and p, where the PDF is given by (1.58) in the following steps:

1. Generate $\boldsymbol{T}_a$, a data set consisted of n_o random numbers t_{a_i} from the exponential distribution with rate parameter a using exprnd$(\frac{1}{a}, n_o, 1)$.

2. Generate $\boldsymbol{T}_b$, a data set consisted of n_o random numbers t_{b_i} from the exponential distribution with rate parameter b using exprnd$(\frac{1}{b}, n_o, 1)$.

3. Generate $\boldsymbol{B}_a$, a data set consisted of n_o random numbers b_{a_i} from the Bernoulli distribution with probability parameter p using binornd$(1, p, n_o, 1)$.

4. Create another data set $\boldsymbol{B}_b = \mathbf{1} - \boldsymbol{B}_a$ consisted of n_o observations b_{b_i}.

5. The ultimate data set $\boldsymbol{T}$, consisted of n_o random numbers t_i from the specified mixture of two exponential distributions, is then given by $b_{a_i}t_{a_i} + b_{b_i}t_{b_i}$.

For a mixture of two geometric distributions with PMF (1.61), a simulated data set $\boldsymbol{N}$ can be obtained in a similar way by following step 1 to 5 and substituting exprnd$(\frac{1}{a}, n_o, 1)$ in step 1 with geornd$(a, n_o, 1)$ and exprnd$(\frac{1}{b}, n_o, 1)$ in step 2 with geornd$(b, n_o, 1)$. In order to obtain robust results, our simulation experiments are based on 10000 replications.

1.8 Outline of Future Chapters

In this chapter we introduced the key statistical distributions studied in the thesis. We also illustrated how mixtures of exponential distributions are linked to mixtures of geometric distributions. According to Theorem 1 and Theorem 2, any method of moments developed for estimating parameters of mixtures of exponentials can be used on the estimation problem of mixtures of geometrics with little amendment. We also introduced a set of data from the experimental study of the prion disease scrapie. Central to the study of prion diseases is the analysis of incubation period data, using the aforementioned distributions. We also propose a new model for the analysis of serial passage experiments from prion biology. In the next chapter we review previous published work on mixture modelling.

The first part of this thesis deals with mixture modelling. Chapter 3 studies four new methods for estimating mixtures of two exponential distributions with modified moments. Chapter 4 applies similar methods to mixtures of two geometric distributions. Our work on mixture modelling is extended to allow negative mixing weight p in Chapter 5. At the end of these three chapters, we compare the performances of all estimators studied to the asymptotically most efficient maximum likelihood estimator.

The second part of this thesis is about the application of our models to real biological data. First, the application of mixture models to prion disease incubation periods (Chapter 6). Second, the development of a novel model, called SRAMPT (Chapter 7). The SRAMPT model provides a new way of estimating sub-clinical infection which cannot be done using experimental procedures alone. In Chapter 8, we study the application of SRAMPT to an epidemiology model based on simulation Markov chains.

Chapter 2

A Review of Literature

Mixture modelling has been the subject of a large and diverse body of literature spanning more than a century. In this chapter, we take a glance at earlier works which have been published on this topic by statisticians and researchers. For a general introduction on finite mixture distributions, applications and detailed description of various statistical methodologies for their analysis, the reader is referred to monographs by Everitt & Hand (1981), Titterington *et al.* (1985) and more recently, by McLachlan & Peel (2000) and Frühwirth-Schnatter (2006).

Pearson (1894) was the originator of the analysis of mixture distributions, whose work has drawn tremendous attention among statisticians for its ability to model heterogeneities in real data. There has been a rapid development in mixture modelling, thanks primarily to the advent of the high speed computer. Throughout, we review key papers on different estimation methods for mixtures of distributions, namely the traditional method of moments, the maximum likelihood estimator, the Bayesian Markov chain Monte Carlo (MCMC) method which is getting more popular, and the informal graphical approaches. We also look at a few papers which discuss the difficult task of determining the number of components of a mixture distribution.

We then turn to the subject of ultimate interest, which is the application of statistical models to real biological data. Our experimental system is that of scrapie disease, the canonical system for the study of the unusual group of prion diseases. We aim to apply our methods to two key features of prion diseases: their incubation periods and the issue of sub-clinical infection. The special role of incubation period in the study of a prion disease which is similar to scrapie is discussed.

2.1 Estimation Methods

The literature surrounding finite mixture distributions is large and many methods have been used for estimating the parameters of such distributions, ranging from informal techniques like the graphical methods to formal methods such as the method of moments, the maximum

likelihood estimators and the Bayesian approaches. In this section, we review this literature and take a look at how these methods can be applied to fitting mixture distributions to sample data.

2.1.1 Graphical Methods

In the past, many authors considered graphical tools and their abilities to detect mixtures in sample. These methods provide us with a quick approach to reveal the mixture before we move on to estimate the parameters. Most of the articles which discussed the use of graphical methods on mixture modelling are based on mixtures of normal distributions.

It seems natural to think that a mixture is present in a sample if its histogram has more than one mode. Haldane (1952) provided a formal test of multimodality, which was questioned by Cox (1966) regarding the sensitivity of the test in regions where the frequencies are low. In fact, we have seen in Chapter 1 that a mixture does not necessarily have more than one mode. Many authors had discovered the conditions for bimodality of a two-component univariate normal mixture (see Harris & Smith (1949), Eisenberger (1964) and Behboodian (1970)). This suggests that a histogram is not a good indicative of a mixture.

Probability plotting is another famous diagnostic tool for the detection of a mixture. The curve of a mixture distribution is somewhat S-shaped, instead of a straight line. In their monograph, Everitt & Hand (1981) presented five probability plots of two-component normal mixture distributions with different degrees of separation between the components. The conclusion was that the departure from linearity is obvious only if the amount of overlap between distributions is small. A somewhat different approach had been attempted by Fowlkes (1979), who modified the probability plots with the objective of "tracking" the presence of a mixture. He compared his methods with other graphical methods, and claimed that his method is more sensitive to the presence of a normal mixture.

A Chi-squared probability plot of the generalised distance of each observation from the sample mean vector can act as an indicative tool of the presence of a mixture in a sample, as discussed in Everitt & Hand (1981). If the data arises from a single distribution, such a plot should be approximately linear. In the case of a mixture the curve will tend to be S-shaped; Everitt and Hand illustrated this behaviour by plotting a Chi-squared probability plot of generalised distances for *iris setosa* and *iris versicolour*, which has a mixture of normal distributions.

Despite their moderate ability to detect mixtures, graphical methods have been widely used to produce initial guesses for parameter values of mixed distributions. Unlike the method of moments and the MLE, one does not have to solve massive arithmetic to derive parameter estimates of a mixture distribution. Hence, graphical methods allow researchers to visually inspect the parameter values of a mixed distribution.

Using a Q-Q plot of a sampled mixture plotted versus the standard normal , Harding (1949) and Cassie (1954) determined the points of inflexion by eye and used them to estimate the mixing weight p of a mixture of normal distributions. Such a visual inspection can

be very subjective and hence it may be difficult to draw a conclusion. A similar approach had been taken by Fowlkes (1979) to estimate p based on fitting a modified logistic curve. However, Fowlkes' method is only plausible when the components are well separated. Bhattacharya (1967) proposed a satisfactory graphical approach for estimating grouped data arising from a m-component normal mixture. By dividing the observations into classes, he showed that a plot of $\log \frac{\phi_{i+1}}{\phi_i}$ versus t_i, where ϕ_i is the observed frequency of class i, and t_i is the mid-point of class i, leads to a series of negative sloped straight lines. He suggested that the estimates of the parameters μ_j and σ_j^2 of the normal mixture can be found using the x-intercept of the line, the class width and the angle between the j^{th} straight line and the negative direction of the x-axis.

We should bear in mind that the best a graphical method can do is provide us with crude parameter estimates. In real life, we may have no *a priori* information about the number of components in a mixture. Therefore, the graphical approach is indeed a helpful tool which draws a picture to tell us the characteristics of a mixture.

2.1.2 The Method of Moments

The application of the method of moments to the problem of estimating the parameters of mixtures of distributions has a long history and dates back to Pearson (1894), Charlier (1906) and Charlier & Wicksell (1924). The method of moments has the attractive property that the moment equations are linear in the mixture proportions. The earliest analysis of mixture distributions was attributed to Pearson (1894) who applied the method of moments to estimate the five parameters in a mixture of two normal distributions . Without a high speed computer, the calculation which involved finding real roots of a polynomial equation of ninth degree was laborious. This polynomial equation is the well known "nonic" equation. Cohen (1967) was among those statisticians who sought to reduce the computational difficulties of solving the nonic equation derived by Pearson (1894). Rider (1961) applied the method of moments to a mixture of two exponential distributions. He mentioned that the moment estimators are consistent as long as the scale parameters of the two components are not identical. Further, he attempted the difficult task of finding the variances of moment estimators when the mixing weight is assumed to be known. Dealing with discrete mixture distributions, Blischke (1962) and Blischke (1964) showed that the asymptotic efficiency of the moment estimators tends to unity as the binomial parameter $n \to \infty$, given identifiability ($n \geq 2m - 1$, where m is the number of components in a mixture). However, when the mixing parameters are known, the asymptotic efficiencies tend to zero.

Although the method of moments is easy and quick to establish, its performance is known to be unsatisfactory. Many users dislike the method because it does not guarantee the estimates of the parameters to be real values and non-negative. The variances of moment estimators can also be extremely large if the separations between the components are narrow. In their paper, Woodward *et al.* (1984) mentioned:

"the method of moments technique often produced unreasonable estimates, such as negative variances, in as many as 25% of realisations from certain mixture of normal models and in more than 60% of realisation from certain mixtures with non-normal components."

Due to these drawbacks, many authors sought to modify the method of moments, with the hope of improving the efficiency of the moment estimator. Joffe (1964) used three half moments when modelling the surface area per cubic centimetre of particles of airborne mine dust with a mixture of exponential densities. Furthermore, he derived the asymptotic variances and covariances of the half moment estimators which lead to the coefficient of variation of the surface area estimate. By resorting the standard moments to fractional moments, Tallis & Light (1968) gained greater efficiency in estimating a mixture of two unknown exponentials. They derived an approximation method to calculate variance of fractional moments estimators, and compared the efficiency, given by the determinant of the asymptotic covariance matrix, relative to the maximum likelihood estimators.

Prior to the advent of computers, the maximum likelihood equations of the mixed Weibull distribution were almost intractable. This is the reason why Falls (1970) attempted the estimation problem of a Weibull mixture model with the traditional method of moments. Using the graphical method of Kao (1959), Falls estimated the mixing probability, p before attempting the estimation problem of the shape parameters and scale parameters of the mixed Weibull distribution. An illustration was made by considering a combined sample of two Weibull distributions. When confronted with more than one set of estimates, Falls adopted Pearson's suggestion by choosing the sets of estimates which produces the closest agreement between the fifth central moment of the sample and the theoretical moment given by the derived estimates. Accordingly, the estimates chosen provided a good fit to the observed data.

John (1970) described the application of the method of moments on mixtures of discrete distributions. Four cases in which both component densities belong to the binomial, the negative binomial, the Poisson or the hypergeometric family were considered. When the sample size is large, Falls found that the asymptotic distribution of the moment estimator was normal. He also compared the method of moments with the maximum likelihood estimator and concluded that the maximum likelihood estimator is somewhat better in identifying the population of origin of observations. Fryer & Robertson (1972) reconsidered the estimation problem given by Pearson (1894) and compared the performance of three techniques, which are the method of moments, the maximum likelihood estimator and the minimum χ^2 estimates. They used Taylor expansions of the moment equations in order to approximate the biases occurring in the five-parameter univariate normal mixture. They concluded there is little difference with regard to bias but for mean square error, the moment estimator is inferior to the the maximum likelihood estimator.Tan & Chang (1972) investigated the efficiency of the method of moments, specifically on the four-parameter normal mixture, and found that the method of maximum likelihood performs better than the method of

moments. They also obtained the asymptotic covariance matrix for the moment estimators. Their results showed that the efficiency of moment estimators is poor, especially when the two component densities are not well separated.

Although method of moments had been replaced by the maximum likelihood estimator and the Bayesian MCMC method, the fact that it provides a quick way to estimate the parameters means that many authors are still interested in this method. Recently, Bening *et al.* (2004) proposed a modified method of moments to build the statistical estimates of the parameters of fractional stable distributions. By using the first three centered logarithmic moment of the data, they found that the offered estimators have high efficiency.

2.1.3 The Maximum Likelihood Estimator

Iterative procedures are vital for the maximum likelihood estimator of mixture distributions. Most of the earlier works on maximum likelihood fitting of mixture distributions are focussed on choosing the most outstanding iterative methods from algorithms such as the steepest descent algorithm and the Newton Raphson (NR) algorithm, until the introduction of the EM algorithm in 1977. Ever since then, there has been a debate on whether the EM algorithm is superior to the NR algorithm in any situation. The convergence of the EM algorithm is guaranteed. However, if converged, the NR is less time consuming than the EM algorithm and is more flexible to allow some degree of extension of the mixture models. Another important issue of the maximum likelihood estimator is its sensitivity to the starting values of the parameters.

Hasselblad (1966) estimated a mixture of m normal distributions with the maximum likelihood estimator using two iterative techniques: the method of steepest descent and the NR procedure. The method of steepest descent appeared better on a small sample while the NR algorithm performed better on a larger sample. More generally, Hasselblad (1969) considered more general random sampling models on mixtures of Poissons, binomials and exponentials in his further paper. He found that starting values are not critical and Aitken's acceleration process always increased the likelihood. Day (1969) showed that the likelihood function of a mixture of normal distributions is unbounded and hence one has to evaluate the likelihood function at each local maximum to determine where the over-all maximum lies. Behboodian (1970) studied the maximum likelihood estimator of a mixture of m normal distributions and found that the maximum likelihood estimates of the over-all mean and variance of the mixture coincide with the sample mean and sample variance. Wolfe (1970) presented cluster analysis of mixtures of multivariate normals and mixtures of multivariate Bernoulli distributions. He agreed that likelihood equations have no closed-form solutions. In the article, he discussed some useful methods for generating initial estimates for mixture analysis. A variety of initial estimates should be tried and the one with the largest likelihood is the best solution for mixture analysis. John (1970) considered the problem of identifying the population of origin of each observation in a sample thought to be drawn from a mixture of two gamma distributions. He applied both the method of moments and the maximum

likelihood method to the solution of the gamma mixture models. Hosmer (1973)'s work on mixture of normal distributions agreed with Hasselblad (1966) and Hasselblad (1969) that initial values does not have much effect on the maximum likelihood estimates.

The pioneering works on the maximum likelihood estimator of mixture distribution were achieved by Dempster *et al.* (1977). They presented an iterative computation of maximum likelihood estimates from incomplete data and named the procedure EM algorithm because every iteration of the algorithm consists of an expectation step followed by a maximisation step. They showed that the likelihood increases monotonically and applied this algorithm on grouped, censored and truncated data. Since then, the EM algorithm became the favourite iterative technique for maximum likelihood estimator. Fowlkes (1979) used a quasi-Newton method to minimise the -log-likelihood function and sum of squares for error of the mixture of two normal distributions. He discussed methods to calculate initial values of parameters by ad hoc methods and found that good starting values are vital to produce accurate estimates.

A survey of literature on the estimation problem of mixture density before 1980's can be found in the Redner & Walker (1984) paper, where they looked at the properties, both theoretical and practical, of the EM algorithm for a mixture of densities from exponential families. Woodward *et al.* (1984) compared the maximum likelihood estimator with the minimum distance estimator of mixing weight and found that the latter does not exhibit the sensitivity to starting values like the former. In Laird *et al.* (1987)'s article which considers the use of the EM algorithm for maximum likelihood estimation of repeated measures data using a multivariate normal model, they suggested a speed up device called Aitken's acceleration to improve the time to convergence of the EM algorithm. A somewhat different approach was taken in Lindstrom & Bates (1988)'s article where they compared the EM, EM with Aitken's acceleration, and NR algorithms as methods for obtaining estimates for the parameters in mixed-effects models for repeated measures data. They sought the best optimisation algorithm which is quick and converges consistently. The conclusion made was that the NR algorithm is preferable in most situations.

To speed up the convergence of the EM algorithm, Böhning *et al.* (1994) took a similar approach to Laird et al by taking Aitken's acceleration-based stopping criterion which is applicable in the case where the log-likelihood sequence is linearly convergent. Focussing on the mixtures of exponential components, Seidel *et al.* (2000a) and Seidel *et al.* (2000b) showed how different starting values and stopping criteria produce different estimates via the EM algorithm. They also studied the power of the likelihood ratio test for exponential homogeneity against mixtures of two exponentials and found that simpler starting strategies (which often fail to approach the global maximum under the null hypothesis) produce better empirical power of the test. In their monograph, McLachlan & Peel (2000) showed that their starting strategy, the use of random starting values, does as well as the deterministic annealing EM (DAEM) algorithm, considered by Ueda & Nakano (1998) in recovering from a poor choice of starting value.

2.1.4 The Bayesian Approach

The initial development of the Bayesian estimator via the Markov Chain Monte Carlo (MCMC) method was slow, and was only made practical when the key paper written by Gelfand & Smith (1990) had been published. The reason for this is that the computation of Bayesian estimators for mixtures of distributions usually leads to intractable calculation, using an older sampling algorithm. Gelfand and Smith reviewed and compared three sampling algorithms to calculate Bayesian posterior densities, which are the stochastic substitution, the Gibbs sampler, and the sampling-importance-resampling algorithm. Before Gelfand and Smith, the MCMC method had been proposed by Tanner & Wong (1987) for the non-parametric estimation of the hazard function from grouped and censored data. In another paper, Diebolt & Robert (1994) showed that the Bayesian sampling, proposed by them, converges to the posterior distribution when dealing with mixture models. They also pointed out that the convergence of the data augmentation algorithm to the true posterior distribution of the parameter is based on a duality principle which requires minimal assumptions about the prior distribution. During the last two decades, the Bayesian MCMC method has become a widely followed approach to the problem of estimating the parameters which determine a mixture density. Escobar & West (1995) studied the application of the Bayesian density estimation of mixtures of normal distributions. In a study of clinical malaria, Vounatsou *et al.* (1998) apply the Bayesian approach to calculate the probabilities of children with different levels of parasitaemia having fever due to malaria, which can be modelled as a mixture of distributions. The estimation of the parameters from a mixture of exponential distributions has been considered by Gruet *et al.* (1999) using the Bayesian framework.

2.2 Determining the Number of Components

When modelling a data set with a mixture distribution, one of the first questions that come into our mind is : "How many components are there in the mixture?". For instance, in cluster analysis, one must know the number of clusters which exist in the data. Unfortunately, the answer to this essential question is not as straightforward as we thought. A natural candidate for testing the number of components in a mixture is the likelihood ratio test. The null hypothesis H_0 states that the number of components in the mixture distribution studied is m_0; whereas the alternative hypothesis H_1 assumes that the number of components is m_1. The test statistic is

$$-2\log\Lambda \tag{2.1}$$

where

$$\Lambda = \frac{L_0}{L_1},$$

L_0 denotes the likelihood function under the H_0, and L_1 represents the likelihood function under H_1. According to Wilks (1938), the asymptotic distribution of the test statistic (2.1), under certain regularity conditions, is a Chi-squared distribution with degree of freedom as the difference between the number of parameters of the two hypotheses. Hasselblad (1969) found satisfactory results when applying the likelihood ratio test on mixtures of exponential, Poisson and binomial distributions. However, many authors (see Wolfe (1970) and Binder (1978)) had made discussion on the irrelevance of the traditional likelihood ratio test for mixture model. The reason for this is, under H_o, the mixing proportions lie on the boundary of the parameter space, and hence the regularity conditions are not fulfilled. As a consequence, the sampling distribution for (2.1) remains as a mystery. Many researchers had worked on obtaining the asymptotic distribution of (2.1); a review of the relevant literature can be found in Everitt & Hand (1981), Section 5.2.2 and Titterington *et al.* (1985), Section 5.4.

Following Aitkin *et al.* (1985), McLachlan (1987) highlighted the role of the bootstrap for the assessment of the null distribution of (2.1) for testing the null hypothesis of a single normal distribution against the alternative hypothesis of a mixture of two normal distributions in the univariate case.

Focusing on the behaviour of the likelihood ratio test, Seidel *et al.* (2000b) found that, for exponential mixture models, a simple starting strategy for the EM algorithm results in a relatively more powerful test when compared to a multiple starting strategy which often come close to the global maximisation of the likelihood.

Based on Kullback & Leibler (1951) information criterion, Vuong (1989) derived a likelihood ratio test and find that the limiting distribution is a weighted sum of independent χ_1^2 random variables when competing models are nested or overlapping; when the competing models are non-nested, the limiting distribution of the test statistics is claimed to be a normal distribution. By extending a theorem by Vuong , Lo *et al.* (2001) demonstrated that the likelihood ratio statistic, based on the Kullback-Leibler information criterion, of the null hypothesis that a random sample is drawn from m_0-component normal mixture distribution against the alternative hypothesis that the sample is arisen from m_1-component normal mixture distribution is asymptotically distributed as a weighted sum of independent Chi-squared random variables with one degree of freedom under general regularity conditions. Their simulation results showed that the test performs well for mixtures of normal distributions with equal variances.

2.3 Disease Incubation Period

In medicine, the incubation period (IP) of infectious disease is defined as the time elapsed from exposure to an infectious agent until clinical signs and symptoms of the disease appear. More precisely, the incubation period is the time required for multiplication of the parasitic organism within the host organism up to the threshold point at which the parasite population

is large enough to produce symptoms in the host (Sartwell (1950)). In epidemiology, a large amount of articles have been published regarding the distribution of incubation period in a variety of infectious diseases. The key paper is attributed to Sartwell (1950) who proposed a robust model which fits the incubation period of common single-exposure infections, using a lognormal distribution ("Sartwell's Law"). Apparently, the model is free from important sources of confounding such as the type of design, age at exposure, measurement errors and population age structure. In a large review of literature, Kondo (1977) investigated the distributions of 82 international infectious disease data sets involving either viruses, bacteria, or parasites, 70% of them followed a lognormal distribution. Many later studies of incubation period, in good agreement with Sartwell's model, follow a lognormal distribution. For instance, neoplastic diseases (Armenian & Lilienfeld (1974)), Mendelian hereditary and metabolic diseases (Armenian & Khoury (1981)), Alzheimer's disease (Horner (1987)), and twinning causative origin (Philippe (1990)). However, there are cases when the lognormal model fails to fit IP distributions; for instance, the incubation period of typhoid fever (Sartwell (1950)); also 30% of the infectious disease data sets studied by Kondo (1977) have IPs which depart from a lognormal distribution due to negative or excessive skewness.

2.3.1 The Importance of the Incubation Period in Prion Research

In Chapter 6, we provide a range of models, including mixture distributions, to describe the incubation period of a prion disease which is similar to scrapie. IP is one of the defining characteristics of TSEs. As described in Chapter 1, the nature of the infectious agent remained a mystery for most of the 20th century, and although now known to involve an infectious protein (prion), there are still uncertainties about the infectious process. In the absence of the infectious agent in most studies, investigators had to rely on two features to characterise the disease: the IP, and patterns of pathology (usually in the brain). Due to its ease of interpretation, and reliability, the use of the IP has dominated the study of prion diseases.

There are many strains of prion diseases such as scrapie (Bruce (1993)). Unlike viruses or bacteria, where strains would be characterised by differences in the parasite genetic sequence, the definition of prion strains is in fact based largely on the characteristics of the incubation period. The IP for a specific strain is "stable" in a given experimental system, but can also be well correlated with features such as pathological damage in areas of the brain, probability of causing infection, and dose-dependent effects. IP are used to identify (if unknown) the dose of prions that an infected animal has received (Prusiner *et al.* (1981)). This method, comparing IP to a standard titration of dose, greatly speeded up research efforts in a field that is hampered by long experiments and resource limitations.

That BSE had caused vCJD in humans was confirmed by experiments which showed that the IP of BSE in a specific rodent model system, was very similar to the IP of vCJD in the same model, and distinct to the IP caused by other prion diseases such as scrapie and sporadic CJD (Bruce *et al.* (1997)). This process is known as "strain typing" by incubation

period.

It is not known how new strains arise, and although strains are stable in experimental systems, they can be induced to change or "mutate" into new strains under certain conditions (or spontaneously). Since, for prion disease, there is no direct analogy to the mutation process of the nucleic acid sequences in parasite genes, there is great interest in characterising the strain process in prion research. The amino acid sequence of PrPc and PrPSc are identical, and it is thought that the strain information must be determined by the structure that the protein folds into. It is sometime suspected that an infection may have been caused by a mixture of strains, and hence the use of mixture distributions to characterise the IP of a prion disease is a promising avenue for exploring the behaviour of strains in a prion infection. This is the topic of Chapter 6.

Chapter 3

Mixtures of Exponential Distributions

Most of the work on mixtures with continuous components which are not normal has been on exponential components. Such mixtures arise in industrial applications, especially in the analysis of failure time data, and have important mathematical properties.

Single exponential distributions are frequently used to model the time interval between successive completely random events (a Poisson process). Exponential models arise in many application, particularly in the analysis of failure data. Examples of variables distributed in this manner would be lifetimes of electronic devices, the gap length between cars crossing an intersection, or arrivals of customers at the check-out counter in a grocery store. Recall from (1.1) that the exponential distribution function is defined as

$$f(t;\theta) = \theta \exp(-\theta t)$$

for $t \geq 0$ and $\theta > 0$.

When one considers that failures may arise for a number of different reasons it seems hardly surprising that a superposition of exponential densities (a mixture) might provide a better description of the failure properties. For example, Davis (1952) fitted an exponential mixture to the failure distribution of electronic valves; Mendenhall & Hader (1958) fitted exponential components to the failure distribution of transmitter receivers.

Generally, a finite exponential mixture has the form of

$$f(t;\Theta) = \sum_{j=1}^{m} p_j \theta_j \exp(-\theta_j t)$$

for $t \geq 0$ where $\Theta = (\boldsymbol{\theta}, p_1, ..., p_{m-1})$ is the vector containing all the unknown parameters in this mixture model and $\boldsymbol{\theta}$ is the vector containing all the rate parameters in $\theta_1, ..., \theta_m$

known *a priori* to be distinct. Also, note that

$$\sum_{j=1}^{m} p_j = 1$$

and

$$\theta_j, p_j > 0$$

for $j = 1, ..., m$.

Let us now recall the clump model in Section 1.7: suppose we have a $m+1$ state Markov process, in which one state is isolated and the rest of the states are clumped into one "level". The generating matrix (as in(1.51)) is in the form of

$$\begin{bmatrix} -\beta & \boldsymbol{u}_c \\ \boldsymbol{v}_c & \boldsymbol{Q} \end{bmatrix}$$

where β is a positive scalar, $\boldsymbol{u}_c$ is a non-negative row m-vector, $\boldsymbol{v}_c$ is a non-negative column m-vector and $\boldsymbol{Q}$ is a $m \times m$ matrix. The lifetime in the level from the time this level is entered into from state 1 to the time of its exit is denoted as T. The survival function of T is given by

$$S(t) = \frac{\boldsymbol{u}_c}{\beta} (\exp \boldsymbol{Q}t) \mathbf{1}_m, \tag{3.1}$$

the PDF of T is

$$f(t) = \frac{\boldsymbol{u}_c}{\beta} (-\boldsymbol{Q}) (\exp \boldsymbol{Q}t) \mathbf{1}_m, \tag{3.2}$$

and its Laplace transform is in the form of

$$L(s) = \frac{\boldsymbol{u}_c}{\beta} \times \frac{-\boldsymbol{Q}}{s\boldsymbol{I} - \boldsymbol{Q}} \times \mathbf{1}_m. \tag{3.3}$$

When $m = 2$, the distribution of T is a linear combination (not necessarily a mixture) of two exponential distributions (note that in this case we may obtain a gamma distribution, see Chapter 5 for details). The parameters of the exponential distributions are the eigenvalues of the matrix $-\boldsymbol{Q}$, which are positive and real if $\boldsymbol{Q}$ is time reversible. In this chapter, we concentrate on the estimation problem of a positive mixture model in which the mixing weights are strictly positive.

Let a denote the rate parameter of the first exponent, b denote the rate parameter of the second exponent and let p be the mixing probability of the first component. If an observation T has a density which can be represented by

$$f(t; a, b, p) = pa \exp(-at) + (1-p) b \exp(-bt) \tag{3.4}$$

where $t \geq 0$, $0 \leq p \leq 1$ and $b > a$ are parameters of the distributions forming the mixture, then we say that T has a mixture of two exponential distributions. For such a mixture

distribution, the mean and variance are given, respectively, by

$$E[T] = p\left(\frac{1}{a}\right) + (1-p)\left(\frac{1}{b}\right)$$

and

$$Var[T] = p\left(\frac{1}{a^2}\right) + (1-p)\left(\frac{1}{b^2}\right) + p(1-p)\left(\frac{1}{a} - \frac{1}{b}\right)^2.$$

The performance of any estimation methods for the analysis of mixed samples depends on the degree of difference between the populations involved. In order to investigate the effect of the separation between the subpopulations on the performance of estimator, we denote r as the ratio of b to a, as in

$$r = \frac{b}{a}.$$

In Figure 3.1, the PDF curve of a two-component exponential mixture, where $r = 2$, is plotted along with the PDF plots of its constituent components. All three curves are quite similar to each other and the intersection takes place at $t = 6.9315$. After the intersection point, the behaviour of all three curves becomes more similar. Figure 3.2 shows the PDF curves for the case when the component densities are well separated ($r = 10$). We observe that the first component has a rather flat PDF curve; the second component has a skewed PDF curve indicating shorter sojourn time t is significantly more probable than longer sojourn time; whereas the PDF curve of the mixture density shows a blend of the characteristics of both components. Before we start the difficult task of parameter estimation for a mixture distribution, we first understand the behaviour of the exponential mixture by drawing the PDF curves of mixtures with different separation. From Figure 3.3, we can see that the PDF curve is more skewed when the separation between the components is larger. In the simulation experiments, we choose different values of r in order to examine how different estimators perform with respect to different separation between the components. Of course, we would expect the estimators to perform better when r is as large as 10 than when the mixture considered has a narrow separation between two constituent components (for example, $r = 2$).

3.1 The Maximum Likelihood Estimator

3.1.1 Introduction

For any distribution, the maximum likelihood estimator (hereafter abbreviated as MLE) returns the estimates of parameters which maximise the likelihood function

$$L(\Theta) = \prod_{i=1}^{n_o} f(t_i; \Theta)$$

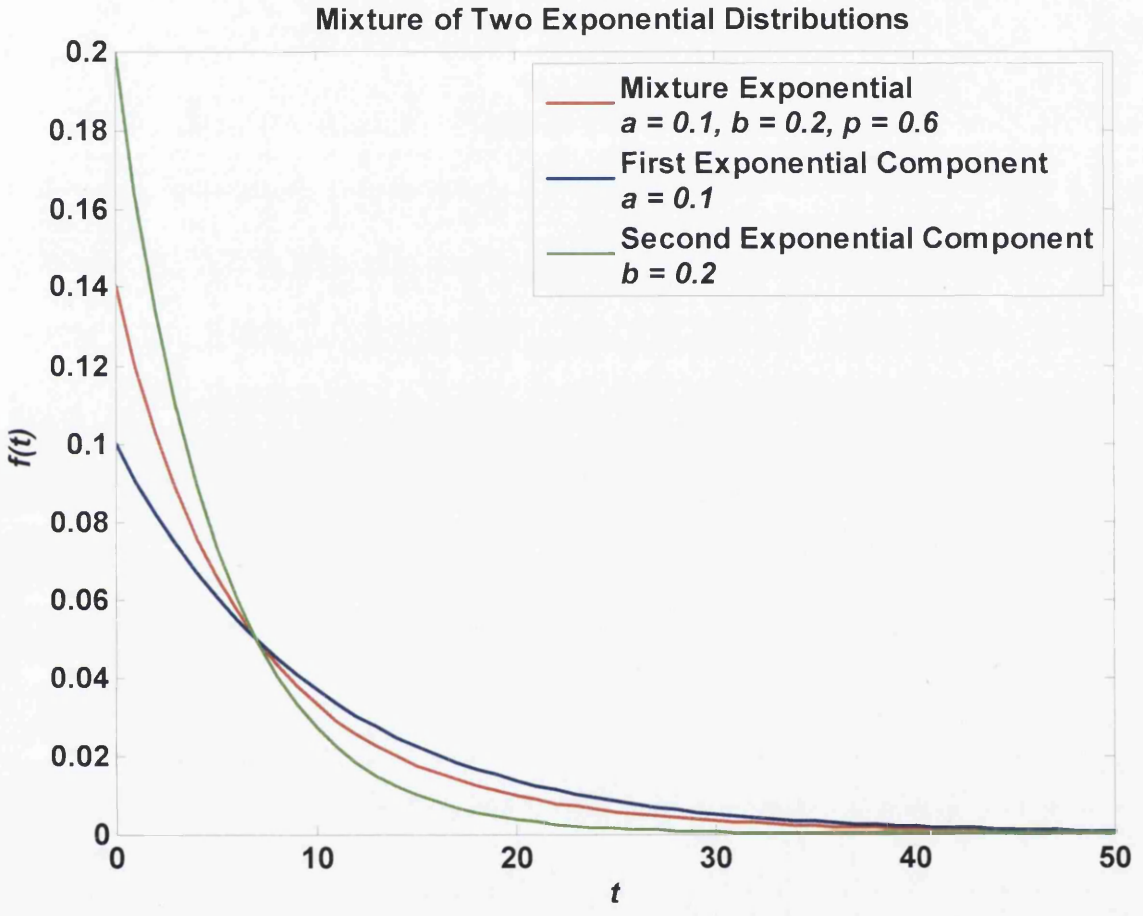


Figure 3.1: PDF plot of a mixture of two exponential distributions together with the PDF plots of its components: $a = 0.1$, $b = 0.2$ and $p = 0.6$.

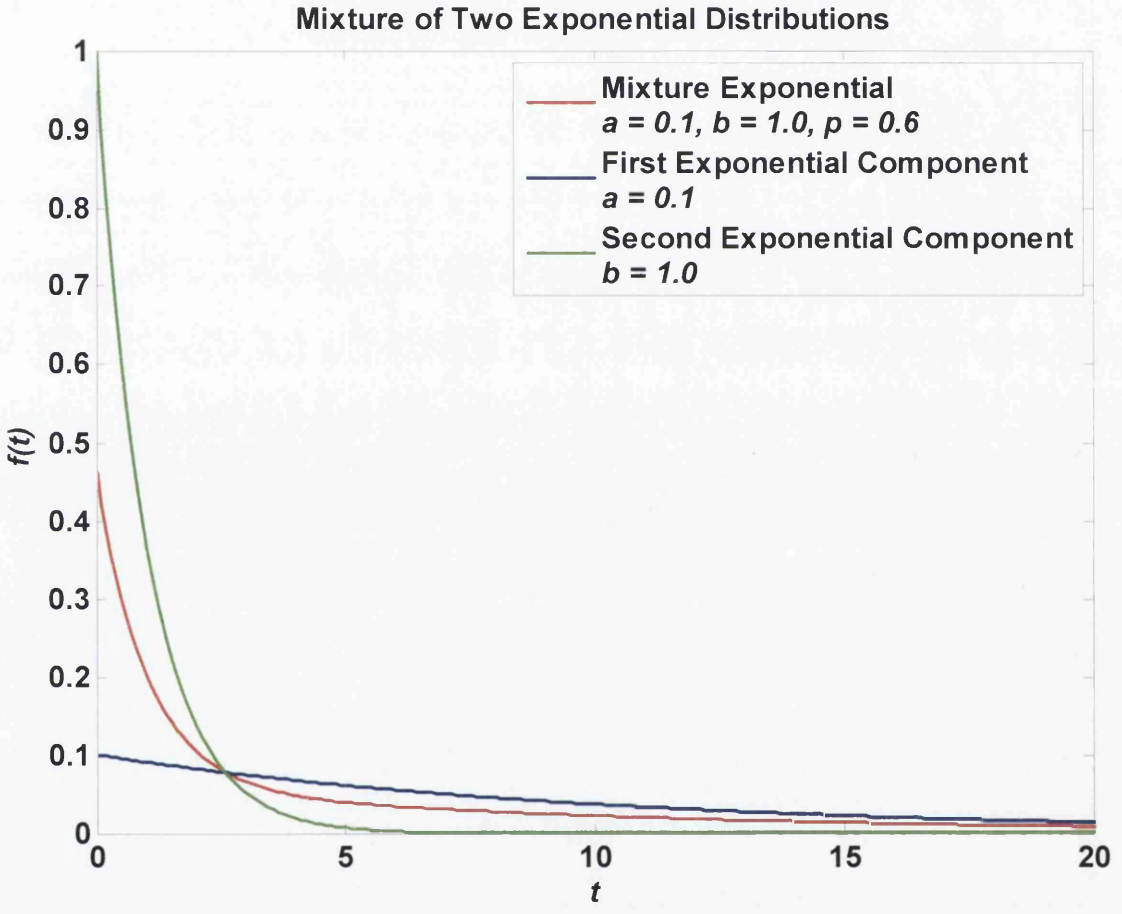


Figure 3.2: PDF plot of a mixture of two exponential distributions together with the PDF plots of its components: $a = 0.1$, $b = 1$ and $p = 0.6$.

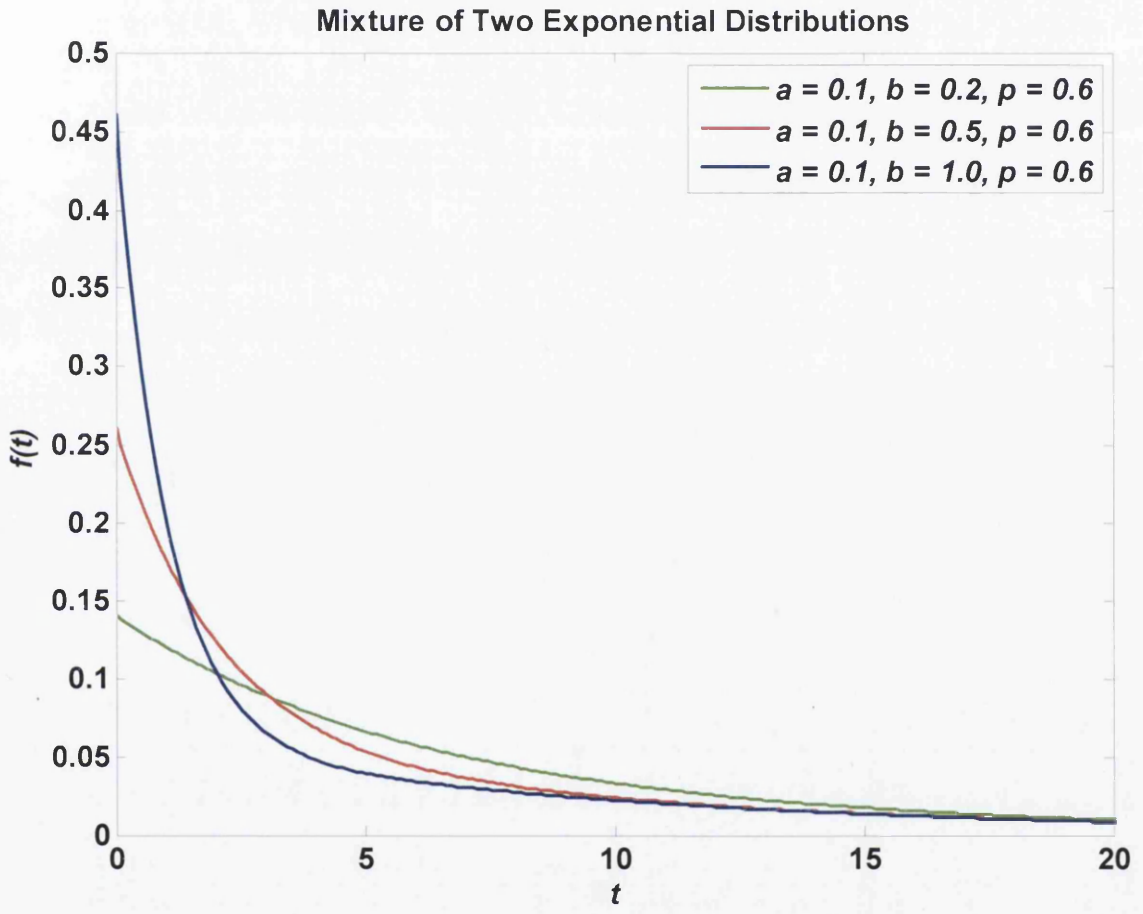


Figure 3.3: PDF plots of mixtures of two exponential distributions for varying separation.

or equivalently, the log-likelihood function

$$l(\boldsymbol{\Theta}) = \log L(\boldsymbol{\Theta}) = \sum_{i=1}^{n_o} \log f(t_i; \boldsymbol{\Theta}).$$

For mixture density problems, the complex dependence of the likelihood function creates a lot of computational difficulties; the likelihood equations for a mixture distribution are almost certain to be nonlinear and there is no way we can find the solution by analytic means. Therefore, iterative methods must be employed in order to produce estimates of a mixture using the MLE. Over the years, there have been great debates, judging from the speed of convergence and the ease of programming, on which iterative procedure outperforms the others. In the following subsections, we would discuss two of the most popular iterative methods, the Newton Raphson's method and the EM algorithm, on the estimation problem of a two-component mixture exponential distribution.

3.1.2 The Newton Raphson's Method

For a mixture of two exponential distributions, we seek the estimates that maximise the log-likelihood function

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{n_o} \log \left[pa \exp(-at_i) + (1-p)\, b \exp(-bt_i) \right]. \tag{3.5}$$

One of the famous approaches is the Newton Raphson's (NR) method which makes use of the score functions and the second derivatives of the log-likelihood function with respect to the parameters. Let $\hat{\boldsymbol{\Theta}}^{(k)}$ denotes the estimate of $\boldsymbol{\Theta}$ $(= (a, b, p))$ obtained after k^{th} iteration of the algorithm. For brevity, we have the following notations

$$l = l(\boldsymbol{\Theta})$$

and

$$\hat{l}^{(k)} = l\left(\hat{\boldsymbol{\Theta}}^{(k)}\right).$$

According to the NR algorithm, the estimates of parameters at the $(k+1)^{th}$ iteration are given by

$$\hat{\boldsymbol{\Theta}}^{(k+1)} = \hat{\boldsymbol{\Theta}}^{(k)} - H\left(\hat{\boldsymbol{\Theta}}^{(k)}\right)^{-1} D\left(\hat{\boldsymbol{\Theta}}^{(k)}\right), \tag{3.6}$$

where $D\left(\hat{\boldsymbol{\Theta}}\right)$ is the Jacobian vector with entries $\dfrac{\partial l}{\partial \hat{\boldsymbol{\Theta}}}$; more specifically,

$$\boldsymbol{D}\left[\hat{\boldsymbol{\Theta}}\right] = \begin{bmatrix} \frac{\partial l}{\partial \hat{a}} \\ \frac{\partial l}{\partial \hat{b}} \\ \frac{\partial l}{\partial \hat{p}} \end{bmatrix} \tag{3.7}$$

where

$$\frac{\partial l}{\partial \hat{a}} = \sum_{i=1}^{n_o} [f(t_i, \hat{\boldsymbol{\Theta}})]^{-1} \left[\hat{p} \exp\left(-\hat{a} t_i\right) - \hat{p} \hat{a} t_i \exp\left(-\hat{a} t_i\right)\right], \tag{3.8}$$

$$\frac{\partial l}{\partial \hat{b}} = \sum_{i=1}^{n_o} \left[f(t_i, \hat{\boldsymbol{\Theta}})\right]^{-1} \left[(1-\hat{p}) \exp\left(-\hat{b} t_i\right) - (1-\hat{p}) \hat{b} t_i \exp\left(-\hat{b} t_i\right)\right], \tag{3.9}$$

$$\frac{\partial l}{\partial \hat{p}} = \sum_{i=1}^{n_o} \left[f(t_i, \hat{\boldsymbol{\Theta}})\right]^{-1} \left[\hat{a} \exp\left(-\hat{a} t_i\right) - \hat{b} \exp\left(-\hat{b} t_i\right)\right], \tag{3.10}$$

and $\boldsymbol{H}\left[\hat{\boldsymbol{\Theta}}\right]$ is the $m \times m$ Hessian matrix

$$\boldsymbol{H}\left[\hat{\boldsymbol{\Theta}}\right] = \begin{bmatrix} \frac{\partial^2 l}{\partial \hat{a}^2} & \frac{\partial^2 l}{\partial \hat{a} \partial \hat{b}} & \frac{\partial^2 l}{\partial \hat{a} \partial \hat{p}} \\ \frac{\partial^2 l}{\partial \hat{a} \partial \hat{b}} & \frac{\partial^2 l}{\partial \hat{b}^2} & \frac{\partial^2 l}{\partial \hat{b} \partial \hat{p}} \\ \frac{\partial^2 l}{\partial \hat{a} \partial \hat{p}} & \frac{\partial^2 l}{\partial \hat{b} \partial \hat{p}} & \frac{\partial^2 l}{\partial \hat{p}^2} \end{bmatrix} \tag{3.11}$$

with entries

$$\frac{\partial^2 l}{\partial \hat{a}^2} = \sum_{i=1}^{n_o} \left[\frac{\hat{p} t_i \left(\hat{a} t_i - 2\right) \exp\left(-\hat{a} t_i\right)}{f\left(t_i\right)} - \left(\frac{\hat{p}\left(1 - \hat{a} t_i\right) \exp\left(-\hat{a} t_i\right)}{f\left(t_i\right)}\right)^2\right],$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \hat{a} \partial \hat{b}} &= \frac{\partial^2 l}{\partial \hat{b} \partial \hat{a}} \\ &= \sum_{i=1}^{n_o} \left[-\frac{\left[\hat{p}\left(1 - \hat{a} t_i\right) \exp\left(-\hat{a} t_i\right)\right] \left[(1-\hat{p}) \left(1 - \hat{b} t_i\right) \exp\left(-\hat{b} t_i\right)\right]}{f\left(t_i\right)^2}\right], \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \hat{a} \partial \hat{p}} &= \frac{\partial^2 l}{\partial \hat{p} \partial \hat{a}} \\ &= \sum_{i=1}^{n_o} \left[\frac{\left(1 - \hat{a} t_i\right) \exp\left(-\hat{a} t_i\right)}{f\left(t_i\right)} - \frac{\left[\hat{p}\left(1 - \hat{a} t_i\right) \exp\left(-\hat{a} t_i\right)\right] \left[\hat{a} \exp\left(-\hat{a} t_i\right) - \hat{b} \exp\left(-\hat{b} t_i\right)\right]}{f\left(t_i\right)^2}\right], \end{aligned}$$

$$\frac{\partial^2 l}{\partial \hat{b}^2} = \sum_{i=1}^{n_o} \left[\frac{(1-\hat{p}) t_i \left(\hat{b} t_i - 2\right) \exp\left(-\hat{b} t_i\right)}{f\left(t_i\right)} - \left(\frac{(1-\hat{p}) \left(1 - \hat{b} t_i\right) \exp\left(-\hat{b} t_i\right)}{f\left(t_i\right)}\right)^2\right],$$

$$\frac{\partial^2 l}{\partial \hat{b} \partial \hat{p}} = \frac{\partial^2 l}{\partial \hat{p} \partial \hat{b}}$$
$$= \sum_{i=1}^{n_o} \left[\frac{\left(\hat{b}t_i - 1\right) \exp\left(-\hat{b}t_i\right)}{f(t_i)} - \frac{\left[(1-\hat{p})\left(1-\hat{b}t_i\right)\exp\left(-\hat{b}t_i\right)\right]\left[\hat{a}\exp\left(-\hat{a}t_i\right) - \hat{b}\exp\left(-\hat{b}t_i\right)\right]}{f(t_i)^2} \right],$$

and

$$\frac{\partial^2 l}{\partial \hat{p}^2} = \sum_{i=1}^{n_o} \left[-\left(\frac{\hat{a}\exp\left(-\hat{a}t_i\right) - \hat{b}\exp\left(-\hat{b}t_i\right)}{f(t_i)} \right)^2 \right].$$

The NR algorithm is well known for its high speed of convergence and flexibility to allow for modified mixture models. Despite its ease of application, many users compared the NR algorithm's performance with other iterative methods and found that it stands out from its competitors in many situations. (Hasselblad (1966), Lindstrom & Bates (1988)). It has also been found to perform better on larger samples (Hasselblad (1966)). However, the main drawback of the NR algorithm is that it does not always converge. It can be problematic if the Hessian matrix in (3.6) is singular and cause the estimates to be undetermined.

3.1.3 The Expectation-Maximisation Algorithm

Dempster *et al.* (1977) introduced a special iterative method called the EM algorithm (E for "expectation" and M for "maximisation"). This procedure is found to have the advantage of reliable global convergence, low cost per iteration, economy of storage and ease of programming. The EM algorithm has been the most popular iterative method for the fitting of mixture distributions with the MLE, thanks to its attractive property of monotonic convergence.

According to the EM algorithm, the observed data is viewed as being incomplete because the component label z_{ij} is missing. The component label tells us which component an observed data t_i comes from:

$$\begin{aligned} z_{ij} &= 1 \qquad \text{if } t_i \sim \exp(\theta_j) \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Therefore, the likelihood function in (3.5) is treated as being incomplete. We assume $z_1, ..., z_{n_o}$, are hidden variables taking values 0 or 1. If z_i takes value 1 then t_i arises from the a-distribution, else we have a b-distribution for t_i. The complete version of the likelihood function is given by

$$l_c(\Theta) = \sum_{i=1}^{n_o} \left[z_{i1} \left(\log p + \log a - a t_i \right) + z_{i2} \left(\log(1-p) + \log b - b t_i \right) \right]. \qquad (3.12)$$

Suppose that $\hat{\Theta}^{(k)}$ denotes the estimate of Θ obtained after k^{th} iteration of the algorithm. At $(k+1)^{th}$ iteration, the E-step handles the addition of the unobservable data to the problem by computing the conditional expected complete data log-likelihood function, denoted by $Q\left(\Theta; \hat{\Theta}^{(k)}\right)$. The M-step maximises $Q\left(\Theta; \hat{\Theta}^{(k)}\right)$ with respect to Θ in order to update the estimate $\hat{\Theta}^{(k)}$. The iteration is normally terminated when the difference between $l\left(\hat{\Theta}^{(k+1)}\right)$ and $l\left(\hat{\Theta}^{(k)}\right)$ is as small as the desired tolerance level. However, other stopping criteria are available and proven to improve the traditional stopping criterion.

E-step

The E-step requires us to calculate the expected value of the log-likelihood over the conditional distribution of the missing data, Z_{ij} given the observed data T and current estimates of parameters $\hat{\Theta}^{(k)}$, which is given by

$$Q\left(\Theta; \hat{\Theta}^{(k)}\right) = E_{\Theta^{(0)}}\left[l_c\left(\Theta\right)|t\right]. \tag{3.13}$$

We know that the complete data log-likelihood l_c is linear in the hidden data z_{ij}. Therefore, to calculate (3.13), one simply needs the conditional expectation of the unobserved data z_{ij}, which is the posterior probability $\tau_j\left(t_i; \hat{\Theta}^{(k)}\right)$ that t_i belongs to the j^{th} component of the mixture

$$\begin{aligned} \tau_j\left(t_i; \hat{\Theta}^{(k)}\right) &= \Pr_{\hat{\Theta}^{(k)}}\left[Z_{ij}=1|t\right] \\ &= \frac{\hat{p}_j^{(k)} f_j(t_i; \hat{\theta}_i^{(k)})}{f(t_i; \hat{\theta}^{(k)})} \\ &= \frac{\hat{p}_j^{(k)} f_j(t_i; \hat{\theta}_j^{(k)})}{\sum_{j=1}^{m} \hat{p}_j^{(k)} f_j(t_i; \hat{\theta}_j^{(k)})}, \end{aligned} \tag{3.14}$$

where $j = 1, ..., m$ and $i = 1, ..., n_o$. Therefore, (3.13) is written as

$$Q\left(\Theta; \hat{\Theta}^{(k)}\right) = \sum_{j=1}^{m}\sum_{j=1}^{n_o} \tau_j\left(t_i; \hat{\Theta}^{(k)}\right)\left[\log p_j + \log f_j\left(t_i; \theta_j\right)\right]. \tag{3.15}$$

For a mixture of two exponential densities, we have

$$\tau_1\left(t_i; \hat{\Theta}^{(k)}\right) = \frac{\hat{p}^{(k)}\hat{a}^{(k)}\exp\left(-\hat{a}^{(k)}t_i\right)}{\hat{p}^{(k)}\hat{a}^{(k)}\exp\left(-\hat{a}^{(k)}t_i\right) + \left(1-\hat{p}^{(k)}\right)\hat{b}^{(k)}\exp\left(-\hat{b}^{(k)}t_i\right)}, \tag{3.16}$$

and

$$\tau_2\left(t_i; \hat{\Theta}^{(k)}\right) = \frac{\left(1-\hat{p}^{(k)}\right)\hat{b}^{(k)}\exp\left(-\hat{b}^{(k)}t_i\right)}{\hat{p}^{(k)}\hat{a}^{(k)}\exp\left(-\hat{a}^{(k)}t_i\right) + \left(1-\hat{p}^{(k)}\right)\hat{b}^{(k)}\exp\left(-\hat{b}^{(k)}t_i\right)}. \tag{3.17}$$

Following (3.15), the conditional expectation required by the E-step is in the form of

$$Q\left(\Theta;\hat{\Theta}^{(k)}\right)=\sum_{i=1}^{n_o}\tau_1\left(t_i;\hat{\Theta}^{(k)}\right)\left[\log p+\log a-at_i\right]+\tau_2\left(t_i;\hat{\Theta}^{(k)}\right)\left[\log\left(1-p\right)+\log b-bt_i\right]. \tag{3.18}$$

M-step

In the EM framework, the updated estimates of mixing weights are obtained independently of the updated estimates of $\boldsymbol{\theta}$. Imagine if z_{ij} are present, the mixing probability p can be estimated easily by

$$\hat{p}=\sum_{i=1}^{n_o}\frac{z_{i1}}{n_o}. \tag{3.19}$$

Replacing z_{i1} in (3.19) with its current expectation $\tau_1\left(t_i;\hat{\Theta}^{(k)}\right)$ (given by (3.16)) the $(k+1)^{th}$ estimate of p is updated as

$$\hat{p}^{(k+1)}=\sum_{i=1}^{n_o}\frac{\tau_1\left(t_i;\hat{\Theta}^{(k)}\right)}{n_o}. \tag{3.20}$$

It should be noted that if the starting value of p is chosen as a positive value, then the updated p is always a positive value. This limits the EM algorithm to give an accurate estimate of p for a linear combination of exponential distributions (where at least one of the p_j's is allowed to be negative). The M-step updates the estimates of $\boldsymbol{\theta}=\{a,b\}$ by globally maximising $Q\left(\Theta;\hat{\Theta}^{(k)}\right)$ with respect to $\boldsymbol{\theta}$ by solving the system of likelihood equations

$$\frac{\partial Q\left(\Theta;\hat{\Theta}^{(k)}\right)}{\partial\boldsymbol{\theta}}=0. \tag{3.21}$$

The $(k+1)^{th}$ updated estimate of $\boldsymbol{\theta}$ is therefore an appropriate root of

$$\sum_{j=1}^{m}\sum_{i=1}^{n_o}\tau_j\left(t_i;\hat{\Theta}^{(k)}\right)\frac{\partial\log f_j\left(t_i;\theta_j\right)}{\partial\boldsymbol{\theta}}=0. \tag{3.22}$$

In order to calculate $\hat{a}^{(k+1)}$, we find the root of the following equation:

$$\sum_{i=1}^{n_o}\tau_1\left(t_i;\hat{\Theta}^{(k)}\right)\frac{\partial}{\partial a}\left(\log a-at_i\right)=0$$

$$\Leftrightarrow\sum_{i=1}^{n_o}\tau_1\left(t_i;\hat{\Theta}^{(k)}\right)\left[\frac{1}{a}-t_i\right]=0.$$

The updated estimates of the rate parameter from the first exponent is

$$\hat{a}^{(k+1)} = \frac{\sum_{i=1}^{n_o} \tau_1\left(t_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)}{\sum_{i=1}^{n_o} t_i \left[\tau_1\left(t_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)\right]}. \tag{3.23}$$

Similarly,

$$\hat{b}^{(k+1)} = \frac{\sum_{i=1}^{n_o} \tau_2\left(t_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)}{\sum_{i=1}^{n_o} t_i \left[\tau_2\left(t_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)\right]}. \tag{3.24}$$

It is clear from (3.16) and (3.17) that $\tau_1\left(t_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)$ and $\tau_2\left(t_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)$ will remain as non-negative if the initial values of $\boldsymbol{\Theta}$ are positive; and hence the updated values of parameters in (3.20), (3.23) and (3.24) will always be non-negative. Therefore, the MLE via the EM algorithm will never return a negative p_j when it is used to estimate the parameters of a sample arisen from a linear combination of exponential distributions (where at least one of the mixing weights is allowed to be negative). In Chapter 5, we illustrate the performance of the MLE in estimating a linear combination of two distributions.

The Stopping Criterion Using Aitken's Acceleration

When should we terminate the EM iteration? Since the incomplète log-likelihood function is increasing monotonically after each EM iteration (as shown by Dempster *et al.* (1977)), a natural candidate for the stopping criterion is when the absolute difference changes by an arbitrarily small amount:

$$\hat{l}^{(k+1)} - \hat{l}^{(k)} < tol \tag{3.25}$$

where *tol* is the tolerance level. A major drawback of the EM algorithm is its slow convergence rate; Lindstrom & Bates (1988) argued that (3.25) is in fact a measure of lack of progress rather than actual convergence.

Laird *et al.* (1987) suggested a useful speeding device called Aitken's acceleration to improve the speed of convergence of the EM algorithm. Böhning *et al.* (1994) also used Aitken's acceleration to reduce the number of iterations for the algorithm. Our investigation also shows that the Aitken's acceleration based stopping criterion does speed up the convergence to the maximum log-likelihood $l_{\max}$. Hence we adopt this stopping criterion to terminate the EM iterative process throughout our simulation experiments in this thesis.

According to Böhning et al., the process assumes that

$$\hat{l}^{(k+1)} - l_{\max} \approx \varepsilon\left(\hat{l}^{(k)} - l_{\max}\right) \tag{3.26}$$

for all k and some ε $(0 < \varepsilon < 1)$, (3.26) is rearranged to give

$$\hat{l}^{(k+1)} - \hat{l}^{(k)} \approx (1-\varepsilon)\left(l_{\max} - \hat{l}^{(k)}\right). \tag{3.27}$$

It is not hard to see that when ε is close to 1, the LHS decreases to a small value. In other words, a small increment in the log-likelihood does not necessarily mean that $\hat{l}^{(k)}$ is approaching the maximum $l_{\max}$. From (3.26), we can see that

$$\hat{l}^{(k+1)} - \hat{l}^{(k)} \approx \varepsilon \left(\hat{l}^{(k)} - \hat{l}^{(k-1)} \right) \tag{3.28}$$

for all k. The limit $l_{\max}$ of the sequence of log-likelihood values $\{\hat{l}^{(k)}\}$ is then obtained using Aitken's acceleration procedure:

$$l_{\max} = \hat{l}^{(k)} + \frac{1}{1-\varepsilon} \left(\hat{l}^{(k+1)} - \hat{l}^{(k)} \right). \tag{3.29}$$

Since ε is unknown, we shall estimate by rearranging (3.28), we estimate the unknown ε by the ratio of successive increments,

$$\hat{\varepsilon}^{(k)} = \frac{\hat{l}^{(k+1)} - \hat{l}^{(k)}}{\hat{l}^{(k)} - \hat{l}^{(k-1)}}, \tag{3.30}$$

which is the ratio of successive increments in log-likelihood. This leads to the Aitken accelerated estimate of $l_{\max}$ (from (3.29)),

$$\hat{l}_A^{(k+1)} = \hat{l}^{(k)} + \frac{1}{(1-\hat{\varepsilon}^{(k)})} (\hat{l}^{(k+1)} - \hat{l}^{(k)}). \tag{3.31}$$

Böhning et al. suggested the EM algorithm can be stopped if

$$\left| \hat{l}_A^{(k+1)} - \hat{l}_A^{(k)} \right| < tol. \tag{3.32}$$

Also, $\hat{l}_A^{(k+1)}$ is then used as a prediction of $l_{\max}$. From (3.31), note that $\hat{l}_A^{(k+1)}$ has a nice monotonicity property when $\hat{\varepsilon}^{(k)}$ is in $(0, 1)$.

3.1.4 Simulation Results

For illustration, a simulation experiment is carried out in order to examine the performance of the MLE for different sample sizes $n_o = (10, 15, 20, 50, 1000)$ and different separations between the two components. All simulations were performed using Matlab. 10000 data sets, each consisting of n_o observations, were simulated from a two-component exponential mixture model with $a = 0.1$, $b = 0.1r$ and $p = 0.6$. The unknown parameter $\boldsymbol{\Theta} = (a, b, p)$ is estimated for each of the 10000 data sets based on the MLE using the EM algorithm. For simplicity, we use the true values of the parameters as the starting values, $\boldsymbol{\Theta}^{(0)} = (0.1, 0.1r, 0.6)$; the stopping criterion adopted is the Aitken's method discussed in (3.32), whereas the tolerance value is set as 0.00001.

Tables 3.1, 3.2 and 3.3 evaluate the estimation error through the average square of bias in estimator $\left(E\left[\hat{\boldsymbol{\Theta}}\right] - \boldsymbol{\Theta} \right)^2$, the variance $Var\left[\hat{\boldsymbol{\Theta}}\right]$ and the mean square error $MSE\left[\hat{\boldsymbol{\Theta}}\right]$ over

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\bar{\hat{a}} - a\right)^2$	0.0005	0.0002	0.0001	1.96×10^{-5}	4.79×10^{-8}
$\left(\bar{\hat{b}} - b\right)^2$	0.3360	0.1931	0.0575	0.0421	0.0014
$\left(\bar{\hat{p}} - p\right)^2$	0.0002	0.0002	0.0004	0.0007	1.80×10^{-5}
$Var\left[\hat{a}\right]$	0.0034	0.0021	0.0017	0.0010	0.0002
$Var\left[\hat{b}\right]$	127	128	3	2	0.0304
$Var\left[\hat{p}\right]$	0.0186	0.0227	0.0261	0.0367	0.0289
$MSE\left[\hat{a}\right]$	0.0040	0.0023	0.0018	0.0010	0.0002
$MSE\left[\hat{b}\right]$	127	128	3	2	0.0319
$MSE\left[\hat{p}\right]$	0.01880	0.0229	0.0265	0.0373	0.0289

Table 3.1: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$ for different sample size n_o. Starting values are set as true values.

all replications for an exponential mixture model with $r = 2$, $r = 5$ and $r = 10$ respectively. For all three separation, both $\left(\bar{\hat{a}} - a\right)^2$ and $Var\left[\hat{a}\right]$ decrease when the sample size increases. However, the ML estimation of b is unsatisfactory for small samples; $Var\left[\hat{b}\right]$ are extremely large for samples with sizes $n_o \leq 50$. When $r = 2$, as seen in Table 3.1, both $\left(\bar{\hat{p}} - p\right)^2$ and $Var\left[\hat{p}\right]$ increase when n_o increases from 10 to 50. Conversely, for large separation $r = 10$, $Var\left[\hat{p}\right]$ decrease when n_o increases.

Figures 3.4, 3.5 and 3.6 show the distribution of the MLEs $\hat{a}$, $\hat{b}$ and $\hat{p}$ for n_o observations arising from a mixture of two exponential distributions with true parameter vector $\boldsymbol{\Theta} = (0.1, 0.5, 0.6)$. In Figure 3.4, the median of $\hat{a}$ is near to the true value $a = 0.1$ for all sample sizes while the number of outliers is greatly reduced when sample size becomes larger. Figure 3.5 allows us to investigate the reason for the large variance of estimator $\hat{b}$ when the sample size is small. As seen from the box plot in the figure, the largest outlier for $n_o = 20$ is $\hat{b} = 8142$, and this explains why $Var\left[\hat{b}\right] = 6815$ in Table 3.2. As seen in Figure 3.6, the estimator $\hat{p}$ has less outliers compared to the rate parameters $\hat{a}$ and $\hat{b}$; $Var\left[\hat{p}\right]$ is considerably large for small samples and is greatly reduced when $n_o = 1000$.

3.1.5 Discussion

To summarise, the EM algorithm works in the following steps:

(a) Set a starting value, $\boldsymbol{\Theta}^{(0)}$.

(b) Set $k = 0$.

(c) E-step: Evaluate $Q\left(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}\right) = E_{\boldsymbol{\Theta}^{(0)}}\left[l_c|t\right]$.

(d) M-step: $\hat{\boldsymbol{\Theta}}^{(k)} = \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}\right)$.

When fitting mixtures of distributions, many users prefer the MLE via the EM algorithm because it is consistent and reliable. The EM algorithm has the advantage of monotonic con-

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	0.0010	0.0003	0.0002	4.87×10^{-6}	3.78×10^{-8}
$(\bar{\hat{b}} - b)^2$	2	3	3	0.1379	7.36×10^{-5}
$(\bar{\hat{p}} - p)^2$	4.87×10^{-6}	4.56×10^{-5}	0.0002	0.0002	1.05×10^{-8}
$Var[\hat{a}]$	0.0073	0.0035	0.0027	0.0009	3.79×10^{-5}
$Var[\hat{b}]$	275	2597	6815	9	0.0043
$Var[\hat{p}]$	0.0409	0.0428	0.0431	0.0358	0.0021
$MSE[\hat{a}]$	0.0083	0.0037	0.0028	0.0009	3.79×10^{-5}
$MSE[\hat{b}]$	277	2600	6818	10	0.0044
$MSE[\hat{p}]$	0.0409	0.0429	0.0433	0.0360	0.0021

Table 3.2: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$ for different sample size n_o. Starting values are set as true values.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	0.0007	0.0002	6.86×10^{-5}	8.34×10^{-6}	1.44×10^{-8}
$(\bar{\hat{b}} - b)^2$	5	4	1.0960	0.1025	3.37×10^{-5}
$(\bar{\hat{p}} - p)^2$	6.09×10^{-6}	9.47×10^{-5}	0.0002	6.23×10^{-5}	4.02×10^{-8}
$Var[\hat{a}]$	0.0084	0.0037	0.0022	0.0007	2.46×10^{-5}
$Var[\hat{b}]$	407	2854	78	11	0.00890
$Var[\hat{p}]$	0.0450	0.0382	0.0348	0.0171	0.0007
$MSE[\hat{a}]$	0.0092	0.0039	0.0023	0.0007	2.46×10^{-5}
$MSE[\hat{b}]$	412	2858	79	11	0.00894
$MSE[\hat{p}]$	0.0450	0.0383	0.0350	0.0172	0.0007

Table 3.3: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$ for different sample size n_o. Starting values are set as true values.

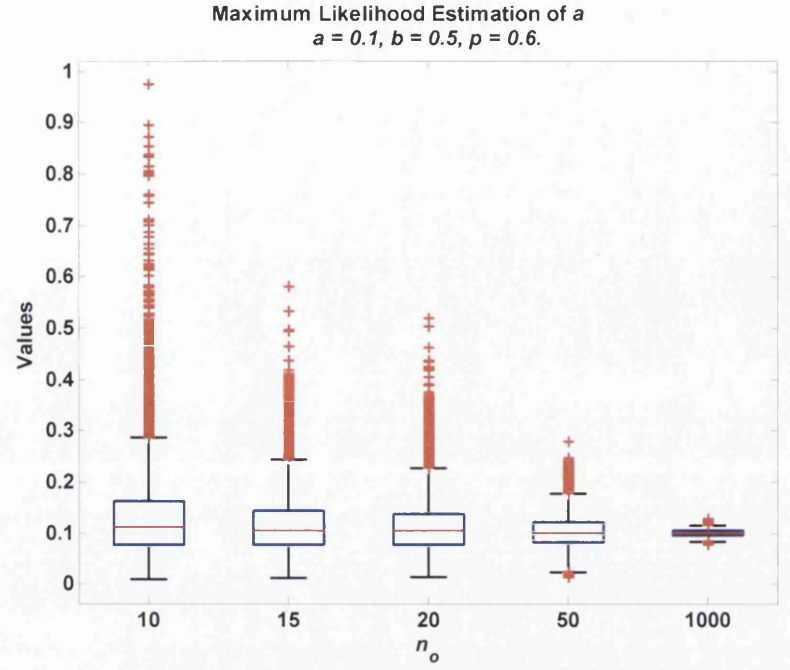


Figure 3.4: Distribution of the MLE $\hat{a}$ for n_o observations arising from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$.

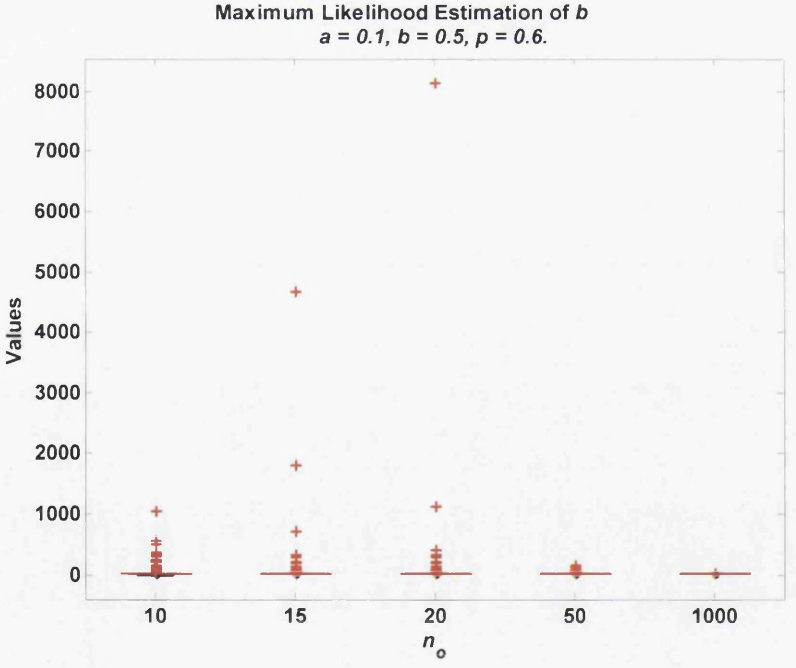


Figure 3.5: Distribution of the MLE $\hat{b}$ for n_o observations arising from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$.

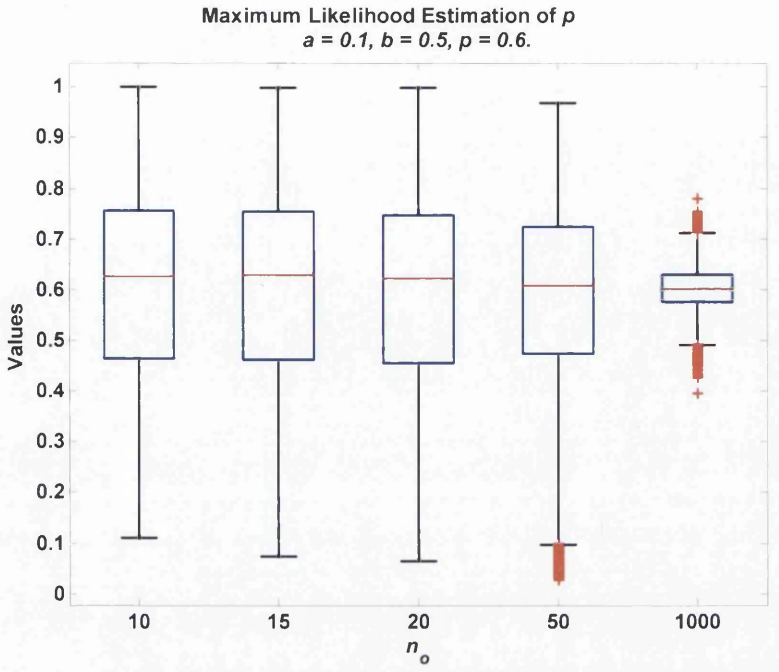


Figure 3.6: Distribution of the MLE $\hat{p}$ for n_o observations arising from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$.

vergence, i.e. the likelihood function increases monotonically with each iteration. However, it might not always be the best method due to some practical difficulties.

The MLE appears to be sensitive to the initial values. The EM algorithm may diverge if the starting values are too close to the boundary of the parameter space. Albert & Baxter (1995) mentioned that the EM algorithm is sensitive to under-specification of the model because the updated value of mixing weight, calculated using (3.20), $p^{(k+1)} = 0$ for all k if $p^{(0)}$ is chosen as 0.

We now study the effect of different starting values on the ML estimates of a mixture of two exponential distributions. Figure 3.7 shows $\hat{a}^{(k)}$, $\hat{b}^{(k)}$, $\hat{p}^{(k)}$ and $\hat{l}^{(k)}$ for each iteration k when we fitted a simulated data set with 1000 observations arising from a two-component exponential mixture distribution with $a = 0.1$, $b = 0.2$ and $p = 0.6$. Starting with $a^{(0)} = 0.1$, $\hat{a}^{(k)}$ deviated from the true value when k increased and it stopped at $\hat{a}^{(36)} = 0.1010$. $\hat{b}^{(1)}$ increased to 0.2004 from the initial true value of b and $\hat{b}^{(k)}$ decreased gradually from $k = 2$ to $k = 13$. After that, $\hat{b}^{(k)}$ increased until the stopping point $k = 36$ with $\hat{b}^{(36)} = 0.2002$; $\hat{b}^{(k)}$ was closest to the true value at $k = 27$ where $\hat{b}^{(27)} = 0.2000$. With the initial point $p^{(0)} = 0.6$, $\hat{p}^{(1)}$ was decreased to 0.5995 and $\hat{p}^{(k)}$ increased gradually after $k = 1$ and the iteration terminated at $\hat{p}^{(36)} = 0.6037$. The log-likelihood of the sample is increased at each iteration and stopped at $\hat{l}^{(36)} = -3070$.

In practice, we will never know the true parameter values of the distribution of a data set, and hence it is worth investigating the performance of the MLE via the EM algorithm

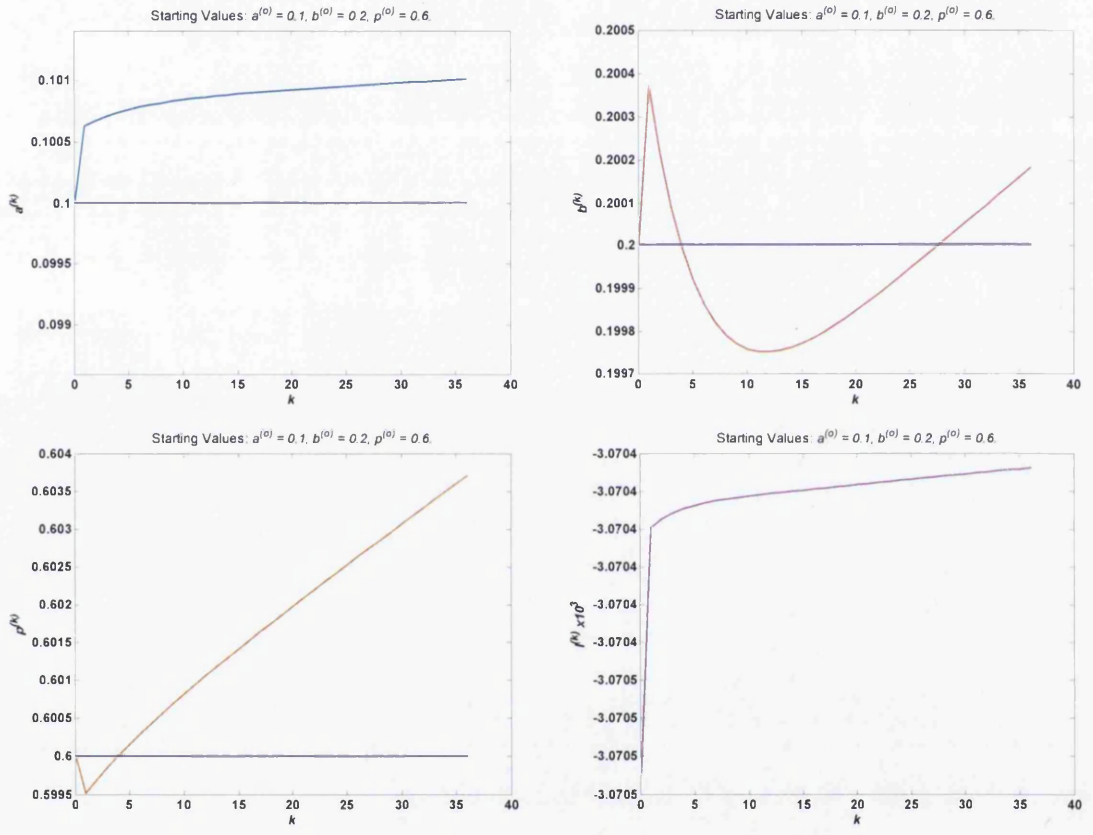


Figure 3.7: The ML updated estimates $\hat{a}^{(k)}$, $\hat{b}^{(k)}$, $\hat{p}^{(k)}$ and $\hat{l}^{(k)}$ at each iteration k for an artificial data set consisting 1000 observations simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$. Starting values are set as true values.

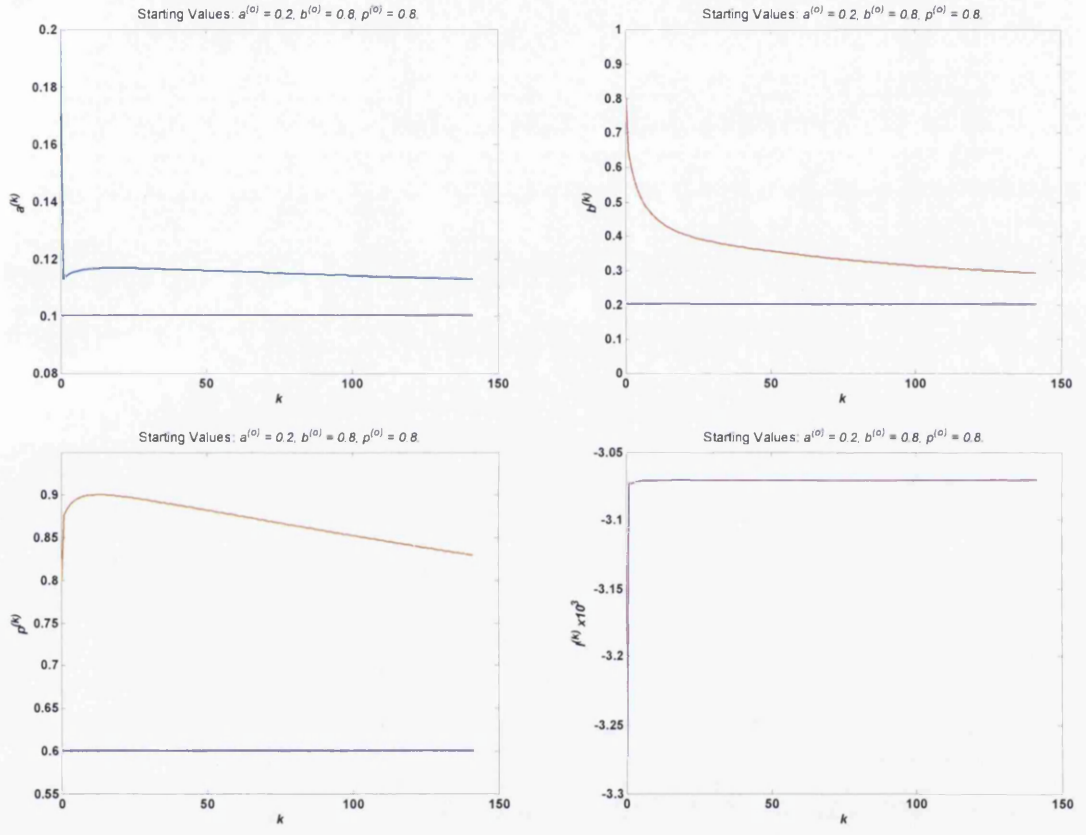


Figure 3.8: The ML updated estimates $\hat{a}^{(k)}$, $\hat{b}^{(k)}$, $\hat{p}^{(k)}$ and $\hat{l}^{(k)}$ at each iteration k for an artificial data set consisting 1000 observations simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$. Starting values are $a^{(0)} = 0.2$, $b^{(0)} = 0.8$ and $p^{(0)} = 0.8$.

when it is used to fit a mixture distribution with starting values differ from the true values of the parameters. On the same data set, we used a different set of starting values $a^{(0)} = 0.2$, $b^{(0)} = 0.8$ and $p^{(0)} = 0.8$ to estimate the parameters with the MLE via the EM algorithm. The EM algorithm terminated after 141 iterations and the resulted estimates are $\hat{a}^{(141)} = 0.1126$, $\hat{b}^{(141)} = 0.2891$, $\hat{p}^{(141)} = 0.8294$ and $\hat{l}^{(141)} = -3071$. The values of the parameters at each iteration k are shown on Figure 3.8. We observed that $a^{(k)}$ decreased sharply at the first iteration from $a^{(0)} = 0.2$ to $\hat{a}^{(1)} = 0.1128$. After this point, $\hat{a}^{(k)}$ increased slowly and had not reached 0.1 even at the 141^{th} iteration. $\hat{b}^{(k)}$ decreased gradually from $b^{(0)} = 0.8$ and failed to arrive at the true value 0.2 when the algorithm stopped. With the initial value $p^{(0)} = 0.8$, $\hat{p}^{(k)}$ first increased gradually to 0.8999 after 13 iterations and decreased slowly after that point with $\hat{p}^{(141)}$ deviating largely from the true value. Studying the plot of $\hat{l}^{(k)}$ versus k, we can see that $\hat{l}^{(1)}$ increased at a large extent to -3074 and the increment was small after this point.

Many authors had pointed out in the past that the likelihood equation usually has a number of local maxima. One would expect the global maximum of the likelihood function to be the "true" root and hence attempt to search all roots of the likelihood function. However, in practice, this attempt could be time consuming and there is no guarantee that the global maximum will be obtained. Furthermore, there are cases when the likelihood function is unbounded and hence the local maximum is mistaken to be the ML estimate.

The EM algorithm has been preferred over the NR algorithm because the time required for each iteration is shorter. The number of iterations required for the NR algorithm is usually small relatively to the number for the EM algorithm. As discussed above, Aitken's acceleration can speed up the EM algorithm but not in all situations (see Lindstrom & Bates (1988)). The EM algorithm will always converge to a local ML function; whereas the NR algorithm does not guarantee convergence. The NR algorithm has an advantage at which the Hessian matrix for the parameter vector is available at the end of the NR iterations. The NR algorithm is superior to the EM algorithm because it can handle most of the common extensions of the mixture models. Lindstrom and Bates preferred the NR algorithm as its advantages outweigh the increase in the computing time per iteration. Although the EM algorithm is a useful tool for ML fitting, there exists potential problems when the number of components are not known *a priori* and the sample size is small.

3.1.6 Information Matrix and Asymptotic Covariance Matrix of the Maximum Likelihood Estimator

The Fisher information is a measure of the amount of information about an unknown parameter contained in an observation. The Fisher information matrix, usually denoted by $\boldsymbol{I}[\boldsymbol{\Theta}]$ plays an important role in the study of the asymptotic properties of MLEs. Given a

sample with n_o observations, the Fisher information matrix can be written as

$$\begin{aligned} n_o \boldsymbol{I}[\boldsymbol{\Theta}] &= n_o E\left[\boldsymbol{\nabla}[\boldsymbol{\Theta}]\boldsymbol{\nabla}[\boldsymbol{\Theta}]^T\right] \\ &= n_o [I_{ij}] \end{aligned} \tag{3.33}$$

where $\boldsymbol{\nabla}[\boldsymbol{\Theta}]$ is the gradient vector with entries $\frac{\partial l}{\partial \boldsymbol{\Theta}}$, which are the score functions. Therefore, $I_{ij} = E\left[\frac{\partial l}{\partial \boldsymbol{\Theta}_i} \cdot \frac{\partial l}{\partial \boldsymbol{\Theta}_j}\right]$. The expected values of the product of score functions for a mixture of two exponential distributions are expressed in the followings:

$$\begin{aligned} I_{aa} &= E\left[\left(\frac{\partial l}{\partial a}\right)^2\right] \\ &= E\left[\left(\frac{p(1-at)\exp(-at)}{f(t)}\right)^2\right] \\ &= \int_0^\infty \frac{(p(1-at)\exp(-at))^2}{f(t)} dt, \end{aligned} \tag{3.34}$$

$$\begin{aligned} I_{ab} &= E\left[\left(\frac{\partial l}{\partial a}\right)\left(\frac{\partial l}{\partial b}\right)\right] \\ &= E\left[\frac{[p(1-at)\exp(-at)][(1-p)(1-bt)\exp(-bt)]}{f(t)^2}\right] \\ &= \int_0^\infty \frac{[p(1-at)\exp(-at)][(1-p)(1-bt)\exp(-bt)]}{f(t)} dt, \end{aligned} \tag{3.35}$$

$$\begin{aligned} I_{ap} &= E\left[\left(\frac{\partial l}{\partial a}\right)\left(\frac{\partial l}{\partial p}\right)\right] \\ &= E\left[\frac{[p(1-at)\exp(-at)][a\exp(-at) - b\exp(-bt)]}{f(t)^2}\right] \\ &= \int_0^\infty \frac{[p(1-at)\exp(-at)][a\exp(-at) - b\exp(-bt)]}{f(t)} dt, \end{aligned} \tag{3.36}$$

$$\begin{aligned} I_{bb} &= E\left[\left(\frac{\partial l}{\partial b}\right)^2\right] \\ &= E\left[\left(\frac{(1-p)(1-bt)\exp(-bt)}{f(t)}\right)^2\right] \\ &= \int_0^\infty \frac{[(1-p)(1-bt)\exp(-bt)]^2}{f(t)} dt, \end{aligned} \tag{3.37}$$

$$\begin{aligned} I_{bp} &= E\left[\left(\frac{\partial l}{\partial b}\right)\left(\frac{\partial l}{\partial p}\right)\right] \\ &= E\left[\frac{[(1-p)(1-bt)\exp(-bt)][a\exp(-at)-b\exp(-bt)]}{f(t)^2}\right] \\ &= \int_0^\infty \frac{[(1-p)(1-bt)\exp(-bt)][a\exp(-at)-b\exp(-bt)]}{f(t)}dt, \end{aligned} \tag{3.38}$$

and

$$\begin{aligned} I_{pp} &= E\left[\left(\frac{\partial l}{\partial p}\right)^2\right] \\ &= E\left[\left(\frac{a\exp(-at)-b\exp(-bt)}{f(t)}\right)^2\right] \\ &= \int_0^\infty \frac{[a\exp(-at)-b\exp(-bt)]^2}{f(t)}dt. \end{aligned} \tag{3.39}$$

No one has successfully expressed the theoretical Fisher information matrix in explicit form. In his notes, Jalali (2008a) found the explicit solution for this difficult task. He expressed (3.34) to (3.39) in terms of three integrals

$$I_0(\varphi, r^*, x) = \int_0^\infty \frac{\exp(-xy)}{1+\varphi\exp(-r^*y)}dy, \tag{3.40}$$

$$I_1(\varphi, r^*, x) = \int_0^\infty \frac{y\exp(-xy)}{1+\varphi\exp(-r^*y)}dy, \tag{3.41}$$

$$I_2(\varphi, r^*, x) = \int_0^\infty \frac{y^2\exp(-xy)}{1+\varphi\exp(-r^*y)}dy, \tag{3.42}$$

where $r^* = r-1$, $\varphi = \dfrac{r(1-p)}{p}$, $y = at$ and x can be any number. Using (3.40) to (3.42), the elements of the Fisher information for a mixture of two exponential distributions ((3.34) to (3.39)) can be written as

$$I_{aa} = \frac{p}{a^2}\left[I_0(\varphi, r^*, 1) - 2I_1(\varphi, r^*, 1) + I_2(\varphi, r^*, 1)\right], \tag{3.43}$$

$$I_{ab} = \frac{1-p}{a^2}\left[I_0(\varphi, r^*, 1+r^*) - (2+r^*)I_1(\varphi, r^*, 1+r^*) + (1+r^*)I_2(\varphi, r^*, 1+r^*)\right], \tag{3.44}$$

$$I_{ap} = -\frac{(1+r^*)}{p}\left[I_0(\varphi, r^*, 1+r^*) - I_1(\varphi, r^*, 1+r^*)\right], \tag{3.45}$$

$$I_{bb} = \frac{(1-p)^2}{a^2 p}\left[I_0(\varphi, r^*, 1+2r^*) - 2(1+r^*)I_1(\varphi, r^*, 1+2r^*) + (1+r^*)^2 I_2(\varphi, r^*, 1+2r^*)\right], \tag{3.46}$$

$$I_{bp} = -\frac{\varphi}{p}\left[I_0\left(\varphi, r^*, 1+2r^*\right) - \left(1+r^*\right) I_1\left(\varphi, r^*, 1+2r^*\right)\right] \tag{3.47}$$

and

$$I_{pp} = \frac{\left(1+r^*\right)^2}{p^3} I_0\left(\varphi, r^*, 1+2r^*\right) - \frac{1}{p^2}. \tag{3.48}$$

Accordingly, there are three versions of the evaluation of the three essential integrals $I_0\left(\varphi, r^*, x\right)$, $I_1\left(\varphi, r^*, x\right)$ and $I_2\left(\varphi, r^*, x\right)$, depending on the value of φ. When $\varphi \leq 1$,

$$I_0\left(\varphi, r^*, x\right) = x^{-1}\ {}_2F_1\left(\gamma, 1; 1+\gamma; -\varphi\right), \tag{3.49}$$

$$I_1\left(\varphi, r^*, x\right) = x^{-2}\ {}_3F_2\left(\gamma, \gamma, 1; 1+\gamma, 1+\gamma; -\varphi\right) \tag{3.50}$$

and

$$I_2\left(\varphi, r^*, x\right) = 2x^{-3}\ {}_4F_3\left(\gamma, \gamma, \gamma, 1; 1+\gamma, 1+\gamma, 1+\gamma; -\varphi\right), \tag{3.51}$$

where $\gamma = \dfrac{x}{r}$. Conversely, when $\varphi > 1$,

$$I_0\left(\varphi, r^*, x\right) = \frac{\pi\varphi^{-\gamma}}{r\sin\left(\pi\gamma\right)} - \frac{\varphi^{-1}}{r\left(1-\gamma\right)}\ {}_2F_1\left(1-\gamma, 1; 2-\gamma; (-\varphi)^{-1}\right), \tag{3.52}$$

$$I_1\left(\varphi, r^*, x\right) = \begin{bmatrix} \dfrac{\pi\varphi^{-\gamma}\left(\pi\cot\left(\pi\gamma\right) + \ln\varphi\right)}{r^2\sin\left(\pi\gamma\right)} \\ +\dfrac{\varphi^{-1}}{r^2\left(1-\gamma\right)^2}\ {}_3F_2\left(1-\gamma, 1-\gamma, 1; 2-\gamma, 2-\gamma; (-\varphi)^{-1}\right) \end{bmatrix} \tag{3.53}$$

and

$$I_2\left(\varphi, r^*, x\right) = \begin{bmatrix} \dfrac{\pi\varphi^{-\gamma}\left(\pi^2\left[1+2\cot^2\left(\pi\gamma\right)\right] + 2\pi\ln\varphi\cot\left(\pi\gamma\right) + \left[\ln\varphi\right]^2\right)}{r^3\sin\left(\pi\gamma\right)} \\ -\dfrac{2\varphi^{-1}}{r^3\left(1-\gamma\right)^3}\ {}_4F_3\left(1-\gamma, 1-\gamma, 1-\gamma, 1; 2-\gamma, 2-\gamma, 2-\gamma; (-\varphi)^{-1}\right) \end{bmatrix}. \tag{3.54}$$

Using Pfaff's transformation, Jalali expressed a single formula for $I_0\left(\varphi, r^*, x\right)$ which encompasses both (3.49) and (3.52):

$$I_0\left(\varphi, r^*, x\right) = x^{-1}\left(1+\varphi\right)^{-\gamma}\ {}_2F_1\left(\gamma, \gamma; 1+\gamma; \frac{\varphi}{1+\varphi}\right) = x^{-1}\left(1+\varphi\right)^{-1}\ {}_2F_1\left(1, 1; 1+\gamma; \frac{\varphi}{1+\varphi}\right). \tag{3.55}$$

He noted the drawback of this special method as the evaluation of $I_1\left(\varphi, r^*, x\right)$ and $I_2\left(\varphi, r^*, x\right)$ are not as straightforward. We validate these formulae by comparing the theoretical values (upper entries) with the observed values (lower entries) of Fisher information in Table 3.4. We find good conformity between the two sets of values. Therefore, one should use (3.43) to (3.48) to find the Fisher information matrix for a mixture of two exponential distributions.

As the regularity conditions hold in the case of mixtures of exponentials, we may also

r	I_{aa}	I_{bb}	I_{pp}	I_{ab}	I_{ap}	I_{bp}
2	49.4087	2.7723	0.3695	6.1756	−3.6798	−0.8768
	(49.3436)	(2.7715)	(0.3695)	(6.1761)	(−3.6781)	(−0.8767)
5	50.7597	0.5208	1.3840	0.0307	−4.7966	−0.5461
	(50.7923)	(0.5207)	(1.3838)	(0.0301)	(−4.7982)	(−0.5459)
10	51.7332	0.1574	2.1780	−0.3845	−4.0052	−0.2903
	(51.7285)	(0.1574)	(2.1781)	(−0.3846)	(−4.0043)	(−0.2904)

Table 3.4: Theoretical (upper) and simulated (lower) Fisher information for a mixture of two exponential distributions with varying r and fixed $a = 0.1$ and $p = 0.6$.

calculate the Fisher information matrix from

$$\begin{aligned} n_o \boldsymbol{I}[\boldsymbol{\Theta}] &= -n_o E[\boldsymbol{H}[\boldsymbol{\Theta}]] \\ &= -n_o \left[I^*_{ij}\right] \end{aligned}$$

where $\boldsymbol{H}[\boldsymbol{\Theta}]$ is the Hessian matrix (as in (3.11)) with entries $\dfrac{\partial^2 l}{\partial \Theta_i \partial \Theta_j}$. In this case, $I^*_{ij} = E\left[\dfrac{\partial^2 l}{\partial \Theta_i \partial \Theta_j}\right]$.

According to the Cramér-Rao bound, the covariance matrix of any unbiased estimator of $\boldsymbol{\Theta}$ is bounded by the inverse of the Fisher information matrix, $[n_o \boldsymbol{I}(\boldsymbol{\Theta})]^{-1}$. We therefore make use of Jalali's solution for the Fisher information matrix to find the Cramér-Rao lower bound (CRLB) of the covariance matrix of estimator for a two-component exponential mixture distribution with fixed $a = 0.1$, $p = 0.6$ and $n_o = 1000$ for three degrees of separation between the two components ($r = (2, 5, 10)$), as shown in Table 3.5. The diagonal elements of the matrices in this table will be used in the last section of this chapter to find the efficiency of all estimators studied by us.

3.1.7 Coincidence of Sample Mean and Theoretical Mean Inferred by the Maximum Likelihood Estimator

In our work regarding the ML estimation of the three parameters of a mixture of two exponential distributions obtained from simulated data, the sample mean of the data appears to be equal to the theoretical mean inferred by the estimated parameters. Let $\hat{T}$ be a non-negative random variable with the mixed exponential PDF

$$\hat{p}\hat{a}\exp(-\hat{a}t) + (1-\hat{p})\,\hat{b}\exp\left(-\hat{b}t\right)$$

where $\hat{a}$, $\hat{b}$ and $\hat{p}$ are obtained from our sample by the MLE, then

$$E\left[\hat{T}\right] = \bar{t}, \tag{3.56}$$

r	CRLB of $\boldsymbol{V}[\boldsymbol{\Theta}]$
2	$\begin{bmatrix} 0.0003 & 0.0011 & 0.0056 \\ 0.0011 & 0.0056 & 0.0243 \\ 0.0054 & 0.0243 & 0.1164 \end{bmatrix}$
5	$\begin{bmatrix} 4.39 \times 10^{-5} & 0.0003 & 0.0003 \\ 0.0003 & 0.0049 & 0.0029 \\ 0.0003 & 0.0029 & 0.0028 \end{bmatrix}$
10	$\begin{bmatrix} 2.68 \times 10^{-5} & 0.0002 & 7.70 \times 10^{-5} \\ 0.0002 & 0.0100 & 0.0017 \\ 7.70 \times 10^{-5} & 0.0017 & 0.0008 \end{bmatrix}$

Table 3.5: Cramér-Rao lower bound of $\boldsymbol{V}[\boldsymbol{\Theta}]$ for a mixture of two exponential distributions with varying r and fixed $a = 0.1$, $p = 0.6$ and $n_o = 1,000$.

where

$$E\left[\hat{T}\right] = \frac{\hat{p}}{\hat{a}} + \frac{1-\hat{p}}{\hat{b}}$$

is the theoretical mean inferred by the MLE and

$$\bar{t} = \sum_{i=1}^{n_o} \frac{t_i}{n_o}.$$

is the sample mean.

Table 3.6 compares the theoretical means inferred by the MLE with the sample means of ten samples drawn from a mixture of exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$, each with sample size, $n_o = 10$. It is obvious that there is a coincidence between the sample mean and the theoretical mean inferred by the MLE.

General Proof Following our conjecture of the coincidence of the sample mean and the MLE inferred theoretical mean of mixture exponentials, Jalali (2008b) investigated the case of an arbitrary distribution from the exponential family and found the following result.

The following is the proof of this phenomenon in the general case of a mixture of m exponential distributions with $2m - 1$ parameters.

Sample	$\hat{a}$	$\hat{b}$	$\hat{p}$	$E\left[\hat{\Theta}\right]$	$\bar{t}$
1	0.1159	2.3837	0.7323	6.4294	6.4294
2	0.0356	0.3308	0.1608	7.0548	7.0548
3	0.1222	1.1947	0.2991	3.0343	3.0343
4	0.1866	1.3475	0.5832	3.4350	3.4350
5	0.0475	0.5967	0.8070	17.307	17.307
6	0.1071	0.9389	0.8195	7.8428	7.8428
7	0.1476	3.7958	0.6765	4.6695	4.6695
8	0.1827	3.4256	0.6047	3.4259	3.4259
9	0.1340	0.2531	0.4745	5.6179	5.6179
10	0.0940	0.5453	0.5439	6.6207	6.6207

Table 3.6: Theoretical means inferred by the MLE ($E\left[\hat{\Theta}\right]$) and sample means ($\bar{t}$) of ten data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$, $p = 0.6$ and $n_o = 1000$.

The Result

Let us have a sample of size n_o of a distribution with the PDF

$$f(t) = \sum_{j=1}^{m} p_j \theta_j \exp(-\theta_j t)$$

where

$$\sum_{j=1}^{m} p_j = 1.$$

The problem is one of the estimation of all the parameters, i.e., p_j's as well as θ_j's.

Theorem 3 *If θ_j's and p_j's are the roots of the maximum likelihood equations, based on the sample $t_1, ..., t_{n_o}$, then*

$$\bar{t} = \frac{1}{n_o} \sum_{i=1}^{n_o} t_i = \sum_{j=1}^{m} \frac{p_j}{\theta_j} = \mu.$$

Proof. *The maximum likelihood equations are of the form:*

$$\frac{\partial l}{\partial p_j} = \sum_{i=1}^{n_o} \frac{\theta_j \exp(-\theta_j t_i) - \theta_m \exp(-\theta_m t_i)}{f(t_i)} = 0, \quad j = 1, ..., m-1 \tag{3.57}$$

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n_o} \frac{p_j \exp(-\theta_j t_i) - p_j \theta_j t_i \exp(-\theta_j t_i)}{f(t_i)} = 0, \quad j = 1, ..., m \tag{3.58}$$

By summing the set of equations (3.58) *over all $j = 1, ..., m$, we obtain the following equation:*

$$\sum_{i=1}^{n_o} \frac{\sum_{j=1}^{m} p_j \exp(-\theta_j t_i)}{f(t_i)} = \sum_{i=1}^{n_o} t_i = n_o \bar{t}. \tag{3.59}$$

We need the following lemma:

Lemma 4

$$\sum_{j=1}^{m-1}\left[p_j\left(\frac{1}{\theta_j}-\mu\right)\right]\left[\theta_j\exp\left(-\theta_j t_i\right)-\theta_m\exp\left(-\theta_m t_i\right)\right]=\sum_{j=1}^{m}p_j\exp\left(-\theta_j t_i\right)-\mu\left[\sum_{j=1}^{m}p_j\theta_j\exp\left(-\theta_j t\right)\right]$$

Proof. *The sum of the LHS can be opened as follows:*

$$\begin{aligned}
LHS &= \sum_{j=1}^{m-1}\left[p_j\left(\frac{1}{\theta_j}-\mu\right)\right]\left[\theta_j\exp\left(-\theta_j t_i\right)-\theta_m\exp\left(-\theta_m t_i\right)\right]\\
&= \sum_{j=1}^{m-1}p_j\exp\left(-\theta_j t_i\right)-p_j\mu\theta_j\exp\left(-\theta_j t_i\right)-\frac{p_j}{\theta_j}\theta_m\exp\left(-\theta_m t_i\right)+p_j\mu\theta_m\exp\left(-\theta_m t_i\right)\\
&= \sum_{j=1}^{m}p_j\exp\left(-\theta_j t_i\right)-p_m\exp\left(-\theta_m t_i\right)-\theta_m\exp\left(-\theta_m t_i\right)\left[\mu-\frac{p_m}{\theta_m}\right]\\
&\quad -\mu\left[\sum_{j=1}^{m}p_j\theta_j\exp\left(-\theta_j t_i\right)-p_m\theta_m\exp\left(-\theta_m t_i\right)-\left(1-p_m\right)\theta_m\exp\left(-\theta_m t_i\right)\right]\\
&= \sum_{j=1}^{m}p_j\exp\left(-\theta_j t_i\right)-\mu\left[\sum_{j=1}^{m}p_j\theta_j\exp\left(-\theta_j t_i\right)\right]\\
&= RHS
\end{aligned}$$

■

■

Proof. *Multiplying each equation in* (3.57) *by* $p_j\left(\frac{1}{\theta_j}-\mu\right)$ *and summing up gives us*

$$\sum_{j=1}^{m-1}\left[p_j\left(\frac{1}{\theta_j}-\mu\right)\right]\sum_{i=1}^{n_o}\frac{\theta_j\exp\left(-\theta_j t_i\right)-\theta_m\exp\left(-\theta_m t_i\right)}{f\left(t_i\right)}=0$$

$$\Rightarrow\sum_{i=1}^{n_o}\frac{\sum_{j=1}^{m-1}\left[p_j\left(\frac{1}{\theta_j}-\mu\right)\right]\left[\theta_j\exp\left(-\theta_j t_i\right)-\theta_m\exp\left(-\theta_m t_i\right)\right]}{f\left(t_i\right)}=0. \qquad (3.60)$$

Now, it follows from the lemma that the term in the inner sum of (3.60) *can be written as*

$$\sum_{j=1}^{m}p_j\exp\left(-\theta_j t_i\right)-\mu\left[\sum_{j=1}^{m}p_j\theta_j\exp\left(-\theta_j t_i\right)\right],$$

(3.60) *then will reduce to*

$$\sum_{i=1}^{n_o} \frac{\sum_{j=1}^{m} p_j \exp(-\theta_j t_i)}{f(t_i)} - n_o \mu = 0.$$

It then follows from (3.59) *that*

$$\frac{1}{n_o}\left(\sum_{i=1}^{n_o} \frac{\sum_{j=1}^{m} p_j \exp(-\theta_j t_i)}{f(t_i)}\right) = \bar{t} = \mu. \tag{3.61}$$

Remark 1 *The proof of the theorem does not depend on the signs of p_j's. So the same theorem is true for a linear combination of exponential distributions.*

Remark 2 *This theorem says that by using the ML method of estimation in mixtures (or linear combinations) of exponentials we also match the first moment of data to the first theoretical moment. So the MLE and the method of moments, as far as the first moment is concerned, agree with one another.*

■

3.2 The Method of Moments

3.2.1 Introduction

The method of moments is a method for constructing estimators of the parameters that is based on matching the sample moments with the corresponding distribution moments. This method consists of evaluating k^{th} sample moments, $\hat{\mu}_k$ and then estimate the parameter, $\boldsymbol{\Theta}$ by solving the equation

$$\hat{\mu}_k = \mu_k(\boldsymbol{\Theta}),$$

where

$$\hat{\mu}_k = \frac{1}{n_o}\sum_{i=1}^{n_o} t_i^k$$

and

$$\mu_k = \int_t t^k f(t)\, dt. \tag{3.62}$$

Pearson (1894) attempted the estimation problem of a normal mixture models with the method of moments and obtained a "nonic" equation. The system of moment equations for a mixed density is generally non-linear. Rider (1961) derived moment estimates for the means in a mixture of two exponential densities under the assumption that the mixing proportions are known.

Let T be a non-negative random variable representing the failure time which has a mixture of two exponential distributions, with density (from (3.4)):

$$f(t) = pa\exp(-at) + (1-p)\,b\exp(-bt).$$

It is easy to see that, using (3.62), the theoretical moments of T are

$$\mu_k = pk!a^{-k} + (1-p)\,k!b^{-k}. \tag{3.63}$$

Let us denote the *normalised* moment as

$$z_k = \frac{\mu_k}{k!},$$

then

$$z_k = pa^{-k} + (1-p)\,b^{-k}. \tag{3.64}$$

We use three of these moments to calculate the parameters and the rest of them for validation. For simplicity, we let $x = a^{-1}$ and $y = b^{-1}$. Hence

$$\begin{aligned} z_1 &= px + (1-p)\,y, \\ z_2 &= px^2 + (1-p)\,y^2, \\ z_3 &= px^3 + (1-p)\,y^2. \end{aligned} \tag{3.65}$$

Rearranging (3.65) yields

$$\begin{aligned} p(x-y) &= z_1 - y, \\ p(x^2-y^2) &= z_2 - y^2, \\ p(x^3-y^3) &= z_3 - y^3. \end{aligned} \tag{3.66}$$

After eliminating p from (3.66) and setting $s = x + y$ and $t = xy$, we have the following values for s and t:

$$\begin{aligned} s &= \frac{z_3 - z_1 z_2}{z_2 - z_1^2}, \\ t &= \frac{z_1 z_3 - z_2^2}{z_2 - z_1^2}. \end{aligned} \tag{3.67}$$

It is clear that x and y are the two roots of the quadratic equation

$$u^2 - su + t = 0 \tag{3.68}$$

(note that x is the larger root and y is the smaller root). So,

$$\begin{aligned} x &= \frac{s+\sqrt{s^2-4t}}{2}, \\ y &= \frac{s-\sqrt{s^2-4t}}{2}. \end{aligned} \tag{3.69}$$

Next, we find the three parameters from

$$\begin{aligned} a &= x^{-1}, \\ b &= y^{-1}, \\ p &= \frac{z_1-y}{x-y}. \end{aligned} \tag{3.70}$$

We obtain the method of moments estimates $\hat{a}$, $\hat{b}$ and $\hat{p}$ by simply replacing z_k's by $\hat{z}_k$'s in (3.67) where the latter are the estimates of moments from the data

$$\hat{z}_k = \frac{1}{n_o k!} \sum_{i=1}^{n_o} t_i^k.$$

To test the validity of the model, we consider one further moment identity, namely,

$$z_4 = p\left(x^4 - y^4\right) + y^4,$$

and substitute for x, y and p from the above ((3.69) and (3.70)). After some routine manipulation, we obtain

$$z_4\left(z_2 - z_1^2\right) - z_3\left(z_3 - z_1 z_2\right) + z_2\left(z_1 z_3 - z_2^2\right) = 0. \tag{3.71}$$

We can use the identity in (3.71) as test statistic for the validity of our model. To be more precise, we define the statistics

$$K_4 = z_4\left(z_2 - z_1^2\right) - z_3\left(z_3 - z_1 z_2\right) + z_2\left(z_1 z_3 - z_2^2\right), \tag{3.72}$$

where z_k's are estimators of *normalised* moments. If this statistic is close to zero, we accept the validity of the model.

To do what we did here more elegantly, and with an eye to generalising it, we construct first the following two matrices:

$$\boldsymbol{M}(u) = \begin{bmatrix} 1 & z_1 & z_2 \\ z_1 & z_2 & z_3 \\ 1 & u & u^2 \end{bmatrix} \tag{3.73}$$

r	2	3	4	5	6
$E[\hat{a}]$	0.1001	0.1020	0.1017	0.1015	0.1017
$E[\hat{b}]$	0.5181	0.3830	0.7102	5	1.3596
$E[\hat{p}]$	0.63655	0.6338	0.6230	0.6185	0.6197
$(\bar{a}-a)^2$	8.80×10^{-9}	3.83×10^{-6}	2.76×10^{-6}	2.19×10^{-6}	2.96×10^{-6}
$(\bar{b}-b)^2$	0.1012	0.0069	0.0962	22	0.5770
$(\bar{p}-p)^2$	0.0013	0.0011	0.0005	0.0003	0.0004
$Var[\hat{a}]$	0.0003	0.0002	0.0002	0.0001	0.0001
$Var[\hat{b}]$	2725	87	328	3.74×10^5	6046
$Var[\hat{p}]$	0.0781	0.0298	0.0195	0.0156	0.0127
$MSE[\hat{a}]$	0.0003	0.0002	0.0002	0.0001	0.0001
$MSE[\hat{b}]$	2725	87	328	3.74×10^5	6047
$MSE[\hat{p}]$	0.0795	0.0310	0.0200	0.0160	0.0130

Table 3.7: Performance of the method of moments for 10000 data sets simulated from a mixture of two exponential distributions with varying $b = 0.1r$ and fixed $a = 0.1$ and $p = 0.6$. r ranging from 2 to 6.

and

$$\boldsymbol{Z}_4 = \begin{bmatrix} 1 & z_1 & z_2 \\ z_1 & z_2 & z_3 \\ z_2 & z_3 & z_4 \end{bmatrix}. \tag{3.74}$$

It can be seen easily that the quadratic equation (3.68) discussed earlier and whose roots were x and y is the following:

$$\det \boldsymbol{M}(u) = 0,$$

and the identity for testing is simply

$$\det \boldsymbol{Z}_4 = 0.$$

Note also that the signed weights can also be calculated as follows:

$$\begin{aligned} \begin{bmatrix} p & 1-p \end{bmatrix} &= \begin{bmatrix} 1 & z_1 \end{bmatrix} \begin{bmatrix} 1 & x \\ 1 & y \end{bmatrix}^{-1} \\ &= -\frac{1}{(x-y)} \begin{bmatrix} 1 & z_1 \end{bmatrix} \begin{bmatrix} y & -x \\ -1 & 1 \end{bmatrix} \\ &= \frac{1}{(x-y)} \begin{bmatrix} z_1 - y & x - z_1 \end{bmatrix}, \end{aligned}$$

which is the same as what we obtained earlier.

r	7	8	9	10
$E[\hat{a}]$	0.1014	0.1011	0.1013	0.1014
$E\left[\hat{b}\right]$	1.8391	1.7911	0.1742	−0.2163
$E[\hat{p}]$	0.61473	0.6115	0.6132	0.6138
$(\bar{\hat{a}} - a)^2$	1.95×10^{-6}	1.26×10^{-6}	1.65×10^{-6}	1.86×10^{-6}
$\left(\bar{\hat{b}} - b\right)^2$	1.2974	0.9823	0.5269	1.4794
$(\bar{\hat{p}} - p)^2$	0.0002	0.0001	0.0002	0.0002
$Var[\hat{a}]$	0.0001	0.0001	0.0001	0.0001
$Var\left[\hat{b}\right]$	4496	12611	6669	6701
$Var[\hat{p}]$	0.0117	0.0112	0.0106	0.0099
$MSE[\hat{a}]$	0.0001	0.0001	0.0001	0.0001
$MSE\left[\hat{b}\right]$	4498	12612	6669	6702
$MSE[\hat{p}]$	0.0119	0.0113	0.0108	0.0101

Table 3.8: Performance of the method of moments for 10000 data sets simulated from a mixture of two exponential distributions with varying $b = 0.1r$ and fixed $a = 0.1$ and $p = 0.6$. r ranging from 7 to 10.

3.2.2 Simulation Results

Tables 3.7 and 3.8 present the simulated results of the moment estimator when the method is applied to mixtures of two exponential distributions. We consider 9 ratios $r\left(=\frac{b}{a}\right)$ ranging from $r = 2$ to $r = 10$, in order to study how the moment estimator behaves on different separations between the two components. For each r, we simulate 10000 data sets each consisting of 1000 observations arising from a mixture of two exponential distributions with parameters $a = 0.1$, $b = 0.1r$, $p = 0.6$. The simulated samples are then fitted with a two component exponential mixture model (as in (3.4)) using the method of moments.

As we expected, when r increases, the mean square errors of $\hat{a}$ and $\hat{p}$ decrease, indicating that the performance of moment estimator is much better for larger separation between the components. However, this is not the case for $\hat{b}$. We observe that the ordinary moment estimator of the second rate parameter has large variance for all r.

We found, from these tables, that $Var\left[\hat{b}\right]$ tends to be larger when r increases. To understand this, let us consider the first three moments of a sample with $r = 10$. The first component has a rate parameter $a = 0.1$, whereas the second component has a rate parameter $b = 1$. The probability that an observation arises from the first component is 0.6. Let us denote z_{ak} as the theoretical k^{th} *normalised* moment of the first exponential component, z_{bk} as the the theoretical k^{th} *normalised* moment of the second exponential component, and z_k denotes the theoretical k^{th} *normalised* moment of the mixture distribution; the theoretical moments of the sample are shown in Table 3.9. We observe that the ratio of the third moments of the two exponents is of an order of 1000. Any effect caused by the smaller moment (from the second component) pales into insignificance compared to the slightest

k	z_{ak}	z_{bk}	z_k	$\frac{z_{ak}}{z_{bk}}$
1	10	1	6.4	10
2	100	1	60.4	100
3	1000	1	600.4	1000

Table 3.9: Theoretical moments z_k of a sample with a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$, and the ratio of the moments of the two exponential components.

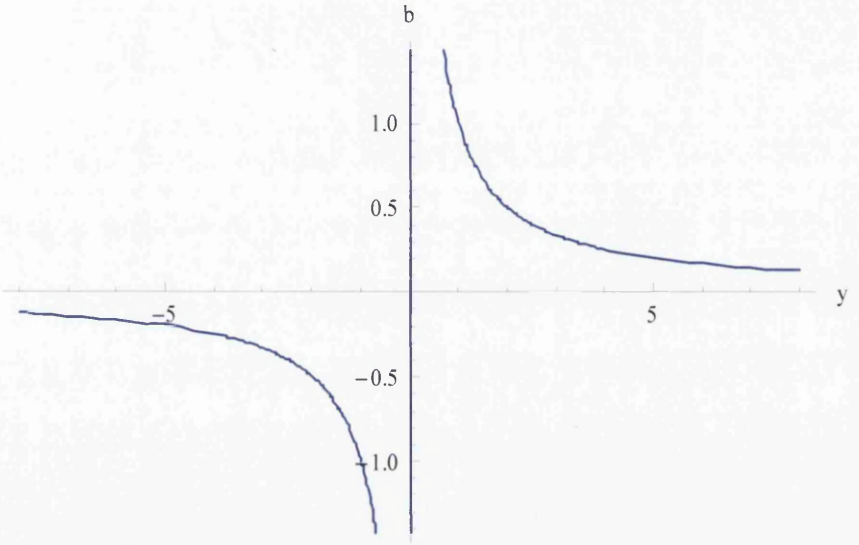


Figure 3.9: Plot of b versus y when $k = 1$.

error in calculation of the larger moment. This is the reason why the estimation of the second component is so poor when the method of moments is adopted.

3.2.3 Discussion

Although the method of moments has long been disfavoured because of its inefficiency relative to the MLE, there is a long history of application of such method because of the computational problems associated with alternative methods, particularly before the advent of computers. There are some potential problems with moment estimators which occur frequently in mixture problem. For example, there is no guarantee that the roots of (3.68) will be real: the method of moments may give estimates of the parameters in complex form. From (3.69), we learn that if $\hat{s}^2 < 4\hat{t}$, $\hat{x}$ and $\hat{y}$ become complex numbers. If $\hat{s}$ is less than $\hat{s}^2 - 4\hat{t}$, then $\hat{y}$ is negative and this makes $\hat{b}$ a negative number. Figure 3.9 is a plot of b against y, given by the formula of b in (3.70). When y has an absolute value close to the origin, b has a large absolute value; the closer is y to the origin, the closer is the absolute value of b approaching to the infinity. Since $\hat{b}$ is very sensitive to $\hat{y}$, the variance of the moment estimator $\hat{b}$ is large, as seen in Table 3.9.

To ensure that $\hat{x}$, $\hat{y} > 0$, we need $\hat{s} > 0$ and $\hat{t} > 0$. In their monograph, Everitt & Hand (1981) stated the conditions on s, t and the three *normalised* moments z_1, z_2 and z_3, as follows:

If we want to know whether the method of moments will work, we should check if

1. $\hat{s}^2 > 4\hat{t}$.
2. The sequence $\frac{1}{z_1}, \frac{z_1}{z_2}, \frac{z_2}{z_3}$ is monotonic.

If we check beforehand that these conditions hold, the moment estimator will be a viable method for the estimation problem of the mixture distribution.

3.3 The Method of Fractional Moments

3.3.1 Introduction

We have seen the method of moments is not very efficient in estimating the parameters of a mixture of two exponential distributions. Now, we will demonstrate how the fractional moments help in controlling the variation between the moments. Our simulation results show that the performance of the moment estimator is greatly improved when the ordinary moments are replaced by the fractional moments. Similar work had been done by Tallis & Light (1968) who reported that efficiencies increase when the method of fractional moments is applied to the estimation problem of a mixture of two exponential distributions.

Jalali (2005c) suggested that the integer k should be replaced by a fraction κ in order to improve the method of moments. The theoretical exposition in this subsection follows his paper. By the moment of order κ of T we mean the integral

$$E\left[T^{\kappa}\right] = \int_0^{\infty} t^{\kappa} f\left(t\right) dt = \kappa \int_0^{\infty} t^{\kappa-1} S\left(t\right) dt, \tag{3.75}$$

where $S\left(T\right)$ is the survival function of T. For the clump model considered in Section 1.7, we know that when $m = 2$, the distribution of the random variable T is a linear combination of two exponential distributions. By substituting the survival function of T, shown in (3.1), into (3.75), the theoretical κ^{th} fractional moment can be written as

$$\begin{aligned} \mu_{\kappa} &= E\left[T^{\kappa}\right] = \kappa \int_0^{\infty} t^{\kappa-1} \frac{\boldsymbol{u}_c}{\mathbf{c}} (\exp \mathbf{Q}t) \mathbf{1}_n dt \\ &= \frac{\kappa \boldsymbol{u}_c}{\mathbf{c}} \left[\int_0^{\infty} t^{\kappa-1} (\exp \mathbf{Q}t) dt \right] \mathbf{1}_n. \end{aligned}$$

Therefore,

$$\mu_{\kappa} = \frac{\Gamma\left(\kappa+1\right) \boldsymbol{u}_c}{\mathbf{c}} \left[-\mathbf{Q}\right]^{-\kappa} \mathbf{1}_n. \tag{3.76}$$

κ	$z_{a\kappa}$	$z_{b\kappa}$	z_κ	$\frac{z_{a\kappa}}{z_{b\kappa}}$
$\frac{1}{3}$	2.1544	1	1.6927	2.1544
$\frac{2}{3}$	4.6416	1	3.1850	4.6416
1	10	1	6.4	10

Table 3.10: Theoretical moments z_κ of a sample with a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$, and the ratio of the moments of the two exponential components.

For simplicity, we use the *normalised* moment

$$z_\kappa = \frac{\mu_\kappa}{\Gamma(\kappa+1)} = \frac{\boldsymbol{u}_c}{\mathbf{c}} [-\mathbf{Q}]^{-\kappa} \mathbf{1}_n. \tag{3.77}$$

In the one dimensional continuous case, T is exponentially distributed. So the survival function is simply

$$S(t) = \exp(-at),$$

and thus

$$\mu_\kappa = \frac{\Gamma(\kappa+1)}{a^\kappa},$$

and as before we denote the *normalised* fractional moment by z_κ, we have,

$$z_\kappa = \frac{1}{a^\kappa}.$$

For a mixture of two exponential distributions, we have the κ^{th} fractional moment

$$\mu_\kappa = \Gamma(\kappa+1)\left[\frac{p}{a^\kappa} + \frac{(1-p)}{b^\kappa}\right], \tag{3.78}$$

and the κ^{th} *normalised* moment

$$z_\kappa = \frac{p}{a^\kappa} + \frac{(1-p)}{b^\kappa}. \tag{3.79}$$

For any positive κ, we will consider the following three moments: z_κ, $z_{\kappa 2}$ and $z_{\kappa 3}$. For simplicity, instead of allowing all three moments to vary, we set $\kappa_2 = 2\kappa_1$ and $\kappa_3 = 3\kappa_1$. From now on, we name κ_1 as κ. We can see from Table 3.10 that the variation between the fractional moments is better controlled, compared to the ordinary moments. The first component has a rate parameter $a = 0.1$, the second component has a rate parameter $b = 1$ and the mixing weight of the first component is 0.6. When $\kappa = 1$, the ratio of the third moment of the first exponent z_{a3} to the second exponent z_{b3} is of an order of 1000 for our example in Table 3.9. However, on the same distribution with the same parameters, the ratio of $z_{a3\kappa}$ to $z_{b3\kappa}$ is largely reduced to 10 when the method of fractional moments (with $\kappa = \frac{1}{3}$) is employed.

If we set $x = a^{-\kappa}$ and $y = b^{-\kappa}$, from our earlier arguments it follows that x and y are

the two roots of the quadratic equation

$$u^2 - su + t = 0, \tag{3.80}$$

where, here,

$$\begin{aligned} s &= \frac{z_{3\kappa} - z_\kappa z_{2\kappa}}{z_{2\kappa} - z_{\kappa^2}}, \\ t &= \frac{z_\kappa z_{3\kappa} - z_{2\kappa^2}}{z_{2\kappa} - z_{\kappa^2}}. \end{aligned} \tag{3.81}$$

From (3.79), p is given by

$$p = \frac{z_\kappa - y}{x - y}. \tag{3.82}$$

If we have a sample $t_1, ..., t_{n_o}$ of size n_o, the estimate of the *normalised* κ-fractional moment is

$$\hat{z}_\kappa = \frac{1}{n_o \Gamma(\kappa + 1)} \sum_{i=1}^{n_o} t_i^\kappa. \tag{3.83}$$

By substituting estimated moments (3.83) for the actual ones in (3.81) the estimates of x and y are given by

$$\hat{x} = \frac{\hat{s} + \sqrt{\hat{s}^2 - 4\hat{t}}}{2} \tag{3.84}$$

and

$$\hat{y} = \frac{\hat{s} - \sqrt{\hat{s}^2 - 4\hat{t}}}{2}. \tag{3.85}$$

The estimate of p can be obtained when we substitute (3.83), (3.84) and (3.85) into (3.82). Obviously, the estimates for parameters a and b are just

$$\begin{aligned} \hat{a} &= \hat{x}^{-\frac{1}{\kappa}}, \\ \hat{b} &= \hat{y}^{-\frac{1}{\kappa}}. \end{aligned} \tag{3.86}$$

3.3.2 Simulation Results

For illustration, a simulation experiment is carried out in order to study the performance of the method of fractional moments for different sample sizes varying from small ($n_o = (10, 15, 20, 50)$) to large ($n_o = 1000$). We consider three degrees of difference between the two populations, ranging from small ($r = 2$) over medium ($r = 5$) to large ($r = 10$). We are interested in two aspects: the minimum variance of estimator and the best fraction κ that returns the smallest measures of error. In order to answer our questions, we consider ten values of fraction: $\kappa = (0.1, 0.2, ..., 1)$. For each κ, we simulated 10000 data sets each consisting of n_o observations from a two-component exponential mixture model with $a = 0.1$, $b = 0.1r$ and $p = 0.6$. We then estimated each data set with the specified κ and recorded the bias^2,

variance and mean square error of the 10000 estimates of $\hat{a}$, $\hat{b}$ and $\hat{p}$. It is known that the method of fractional moments would return complex estimates. During our estimation, we excluded any estimate in complex form to make it easier for us to analyse the results. For each n_o and r, we found the minimum measures of error and present them in Tables 3.11 to 3.13. The counterpart κ's are shown in brackets in these tables.

We observe that, regardless of the separation between the components, the best fraction, in terms of both the bias and variance, for estimating b is 1 when the sample size is small ($n_o \leq 50$). Does this mean that the ordinary moment estimator is better than the fractional moment estimator in estimating b for small samples? We further investigated the estimator and found that more than half of the estimates of b are negative when $\kappa = 1$. For instance, when $r = 2$ and $n_o = 10$, 61.69% of the estimates of b are negative when $\kappa = 1$. When $\hat{b}$ is a large negative number, $\hat{p}$ is close to 1. In other words, the distribution fitted by the method of ordinary moments is actually a single exponential distribution, rather than a mixture. On the other hand, for samples of the same size and same r, when $\kappa = 0.6$, 1.5% of the 10000 estimates of b are over-estimated and they are greater than 100; these outliers make the variance of $\hat{b}$ extremely large when $\kappa = 0.6$. The reasons for these large variations have been investigated and the explanation will be given in Section 3.3.5. Since most of $\hat{b}$'s are negative with a small absolute value when $\kappa = 1$, $E\left[\hat{b}\right]$ is -0.1450 and hence $\left(\overline{\hat{b}} - b\right)^2$ smaller when $\kappa = 1$, compared to the ones given by $\kappa = 0.6$.

For samples of size $n_o = 10$, we found that 75.48% of the ordinary moment estimators $\hat{b}$ are negative when $r = 5$; whereas 83.86% of $\hat{b}$ are negative when $r = 10$. These negative $\hat{b}$'s make both the $\left(\overline{\hat{b}} - b\right)^2$ and $Var\left[\hat{b}\right]$ relatively smaller compared to the fractional moment estimator. In fact, the estimates of b given by the method of fractional moments are more reasonable than the ordinary moment estimators: at least it recognises most of the distribution as a positive mixture of exponential distributions for small samples.

Let us now focus on the large samples ($n_o = 1000$) and study the performance of the fractional moment estimator. For small separation ($r = 2$), as seen in Table 3.11, the best fraction for estimating a is 0.9 because it has the smallest variance and mean square error. For b, the optimal κ is 0.6 as it has both the smallest bias and variance. Compared to the ordinary moment estimator where $Var\left[\hat{b}\right] = 2725$, the fractional moment estimator has successfully reduced the variance of $\hat{b}$ to 14. Both the bias and variance of $\hat{p}$ is minimised when $\kappa = 0.9$.

It is obvious from Table 3.12 that the best fraction for estimating all three parameters (a, b and p) is 0.4 when the ratio of $r = 5$. It is worth noting that, the fractional moment estimator has significantly improved the estimation of b and $Var\left[\hat{b}\right]$ is 0.0079 when $\kappa = 0.4$ (compared to $Var\left[\hat{b}\right] = 3.74 \times 10^5$ when $\kappa = 1$, as shown in Table 3.7). With this fraction, the estimates are both lowly biased and have small variances.

As seen in Table 3.13, $\kappa = 0.3$ best estimates all three parameters when the two populations have a large separation ($r = 10$). We are pleased to see the variance of $\hat{b}$ reduced to

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	1.39×10^{-9} (0.6)	6.99×10^{-8} (0.4)	4.46×10^{-8} (0.3)	3.15×10^{-7} (0.2)	8.80×10^{-9} (1)
$(\hat{b} - b)^2$	0.1190 (1)	0.0051 (1)	0.4632 (1)	0.1369 (1)	0.0592 (0.6)
$(\hat{p} - p)^2$	1.22×10^{-5} (0.8)	0.0016 (0.6)	7.70×10^{-5} (0.7)	0.0003 (0.4)	0.0001 (0.9)
$Var[\hat{a}]$	0.0039 (1)	0.0028 (1)	0.0023 (1)	0.0015 (1)	0.0003 (0.9)
$Var[\hat{b}]$	138 (1)	59 (1)	647 (0.9)	316 (1)	14 (0.6)
$Var[\hat{p}]$	0.4859 (0.2)	0.3631 (0.1)	0.3543 (0.2)	0.2038 (0.2)	0.0702 (0.9)
$MSE[\hat{a}]$	0.0039 (1)	0.0028 (1)	0.0023 (1)	0.0016 (1)	0.0003 (0.9)
$MSE[\hat{b}]$	138 (1)	59 (1)	648 (0.9)	317 (1)	14 (0.6)
$MSE[\hat{p}]$	0.5612 (0.2)	0.4313 (0.1)	0.4009 (0.2)	0.2234 (0.2)	0.0704 (0.9)

Table 3.11: Performance of the method of fractional moments for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	5.16×10^{-6} (0.9)	2.62×10^{-7} (0.8)	8.39×10^{-8} (0.3)	1.28×10^{-7} (0.8)	2.41×10^{-8} (0.3)
$(\hat{b} - b)^2$	0.2758 (1)	0.0943 (1)	0.1891 (1)	5 (1)	0.0002 (0.3)
$(\hat{p} - p)^2$	0.0004 (0.6)	3.09×10^{-6} (0.3)	2.82×10^{-5} (0.2)	9.36×10^{-6} (0.7)	9.55×10^{-8} (0.4)
$Var[\hat{a}]$	0.0064 (1)	0.0037 (1)	0.0027 (0.9)	0.0010 (0.8)	5.27×10^{-5} (0.4)
$Var[\hat{b}]$	21559 (1)	4082 (1)	152 (1)	3014 (0.9)	0.0079 (0.4)
$Var[\hat{p}]$	0.2073 (1)	0.1801 (1)	0.1300 (0.4)	0.0503 (0.8)	0.0037 (0.4)
$MSE[\hat{a}]$	0.0069 (0.8)	0.0039 (0.8)	0.0028 (0.6)	0.0010 (0.8)	5.27×10^{-5} (0.4)
$MSE[\hat{b}]$	21559 (1)	4082 (1)	152 (1)	3024 (0.9)	0.0082 (0.4)
$MSE[\hat{p}]$	0.2188 (1)	0.1882 (0.3)	0.1301 (0.4)	0.0509 (0.8)	0.0037 (0.4)

Table 3.12: Performance of the method of fractional moments for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	4.74×10^{-5} (0.9)	7.32×10^{-6} (0.9)	3.05×10^{-6} (0.2)	2.11×10^{-9} (0.2)	1.60×10^{-9} (0.2)
$(\bar{\hat{b}} - b)^2$	0.5748 (1)	110 (0.9)	0.1067 (1)	0.1332 (1)	0.0001 (0.2)
$(\bar{\hat{p}} - p)^2$	1.17×10^{-5} (0.3)	7.68×10^{-5} (0.4)	0.00011 (0.3)	1.07×10^{-7} (0.3)	5.89×10^{-8} (0.3)
$Var[\hat{a}]$	0.0094 (1)	0.0039 (0.6)	0.0025 (0.4)	0.0007 (0.3)	3.21×10^{-5} (0.3)
$Var[\hat{b}]$	549 (1)	54539 (0.9)	1483 (1)	18 (0.2)	0.01765 (0.3)
$Var[\hat{p}]$	0.1246 (0.4)	0.0651 (0.6)	0.0506 (0.7)	0.0235 (0.4)	0.0012 (0.3)
$MSE[\hat{a}]$	0.0103 (0.6)	0.0041 (0.3)	0.0026 (0.4)	0.0008 (0.3)	3.22×10^{-5} (0.3)
$MSE[\hat{b}]$	550 (1)	54649 (0.9)	1483 (1)	20 (0.2)	0.0179 (0.3)
$MSE[\hat{p}]$	0.1247 (0.4)	0.0659 (0.6)	0.0525 (0.7)	0.0237 (0.3)	0.0012 (0.3)

Table 3.13: Performance of the method of fractional moments for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$ for different sample size n_o.

0.1765 when $\kappa = 0.3$ (recall that the variance of $\hat{b}$ is 6701 when $\kappa = 1$).

Judging from the estimation results of large samples, we should use a large fraction to estimate the parameters when r is small, for example, $\kappa = 0.6$ when $r = 2$; whereas for distribution with well separated components, we should use small fraction, for instance, $\kappa = 0.3$ when $r = 10$.

3.3.3 Asymptotic Covariance Matrix of the Fractional Moment Estimator

We now devise the asymptotic covariance matrix of the fractional moment estimator following the procedures described in Section 1.5.5. A similar approach had been taken by Tallis & Light (1968). The large sample variance of the fractional moment estimator $\boldsymbol{\Theta}$ can be approximated by

$$\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right] \approx \boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1} \boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right] \left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1}\right)^{T}, \tag{3.87}$$

r	Optimal κ	$Var[\hat{a}]$
2	0.5910	0.0003
3	0.4553	0.0001
4	0.3893	6.78×10^{-5}
5	0.3491	5.22×10^{-5}
6	0.3217	4.44×10^{-5}
7	0.3016	3.92×10^{-5}
8	0.2860	3.59×10^{-5}
9	0.2735	3.35×10^{-5}
10	0.2633	3.16×10^{-5}

Table 3.14: Optimal fraction κ and theoretical minimum variance of the fractional moment estimator $\hat{a}$ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

where $\boldsymbol{D}[\boldsymbol{\Theta}]$ has entries $\dfrac{\partial \mu_{\kappa_i}}{\partial \boldsymbol{\Theta}_j}$, with

$$\begin{aligned} \frac{\partial \mu_{\kappa_i}}{\partial a} &= -\Gamma(\kappa_i + 1)\frac{p\kappa}{a^{\kappa_i+1}}, \\ \frac{\partial \mu_{\kappa_i}}{\partial b} &= -\Gamma(\kappa_i + 1)\frac{(1-p)\kappa}{b^{\kappa_i+1}}, \\ \frac{\partial \mu_{\kappa_i}}{\partial p} &= \Gamma(\kappa_i + 1)\left[\frac{1}{a^{\kappa_i}} - \frac{1}{b^{\kappa_i}}\right], \end{aligned} \tag{3.88}$$

for $i = 1, 2, 3$. For simplicity, we set κ_2 as 2κ and κ_3 and 3κ for our investigation. $\boldsymbol{V}[\hat{\boldsymbol{\mu}}]$ is the covariance matrix of the sample fractional moments with entries

$$V_{ij}[\hat{\boldsymbol{\mu}}] = Cov[\hat{\mu}_i, \hat{\mu}_j] = \frac{\hat{\mu}_{i+j} - \hat{\mu}_i\hat{\mu}_j}{n_o}, \tag{3.89}$$

where $\hat{\mu}_i$ is in the form of (3.78).

3.3.4 Optimal Fraction κ

From (3.87), we obtain an approximated variance of fractional moment estimators for a, b and p. Since the variances can be expressed as functions of a, b, p, κ and n_o, it is possible for us to find the optimal fraction κ which minimises the variances given the values of a, b, p and n_o. The variances of the estimators are the diagonal elements of $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ in (3.87). Using the built-in function "FindMinimum" in Mathematica, we obtain the value of κ which minimises $Var[\hat{a}]$, $Var\left[\hat{b}\right]$ and $Var[\hat{p}]$ respectively. Tables 3.14 to 3.16 present the best value of κ for the parameter estimation of mixture exponential distributions with true parameters $a = 0.1$, $b = 0.1r$ and $p = 0.6$. We consider different degree of separation between the two components ranging from small ($r = 2$) to large ($r = 10$).

Figure 3.10 shows nine plots of $Var[\hat{a}]$ against κ; each plot represents different degree of separation between the two components. Table 3.14 presents the optimal fraction and the

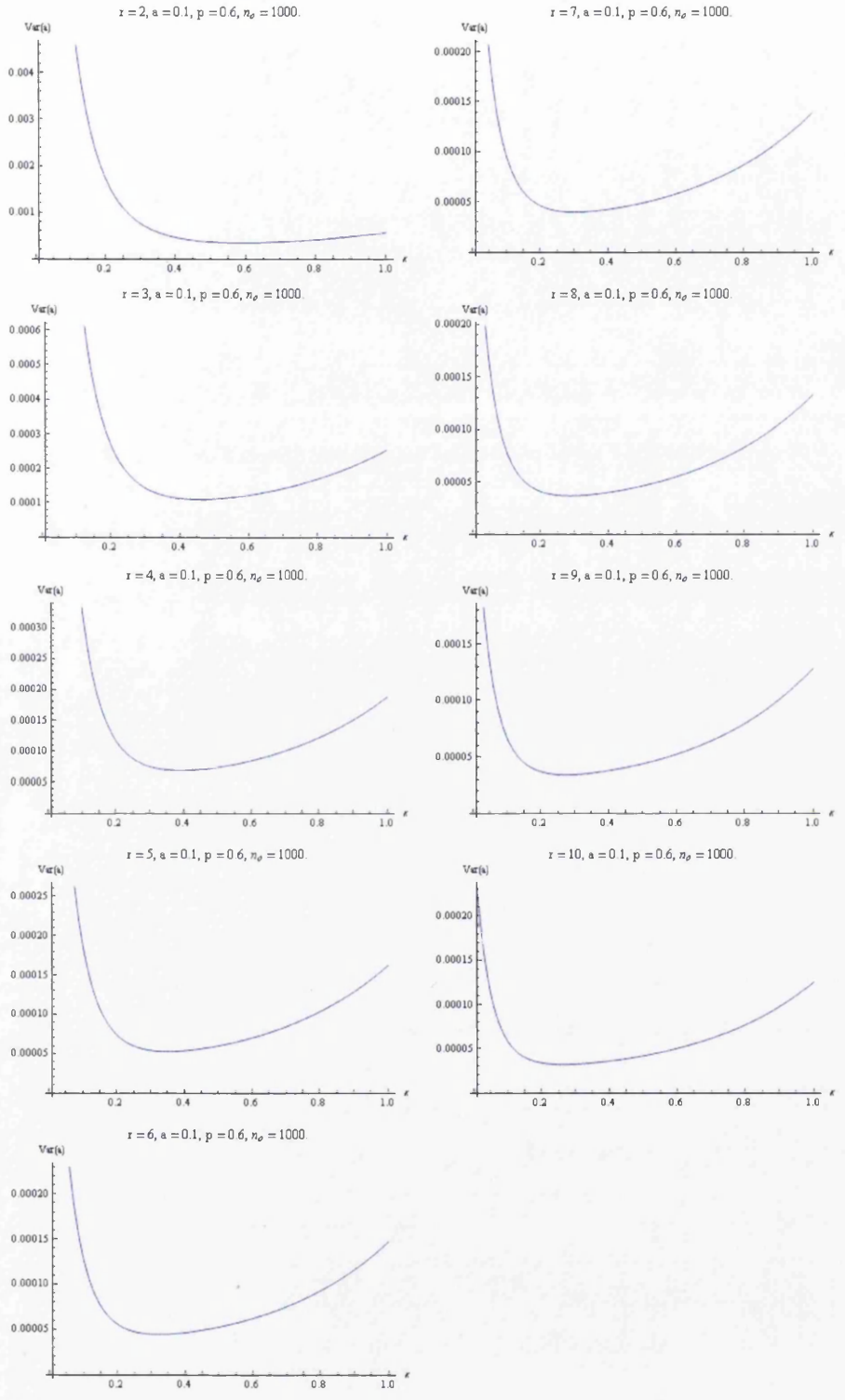


Figure 3.10: Plots of $Var\left[\hat{a}\right]$ versus κ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

r	Optimal κ	$Var[\hat{b}]$
2	0.6262	0.0063
3	0.5010	0.0051
4	0.4378	0.0059
5	0.3983	0.0071
6	0.3708	0.0086
7	0.3501	0.0104
8	0.3339	0.0124
9	0.3208	0.0146
10	0.3098	0.0169

Table 3.15: Optimal fraction κ and theoretical minimum variance of the fractional moment estimator $\hat{b}$ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

r	Optimal κ	$Var[\hat{p}]$
2	0.6064	0.1318
3	0.4727	0.0167
4	0.4056	0.0065
5	0.3637	0.0037
6	0.3345	0.0025
7	0.3126	0.0019
8	0.2955	0.0015
9	0.2815	0.0013
10	0.2699	0.0012

Table 3.16: Optimal fraction κ and theoretical minimum variance of the fractional moment estimator $\hat{p}$ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

minimum variance of $\hat{a}$ for each r considered. When $r = 2$, the minimum point of $Var[\hat{a}]$ occurs at $\kappa = 0.5910$; whereas the best fraction which minimises $Var[\hat{a}]$ when $r = 5$ is 0.3491. As the separation between the two components increases, both the optimal κ and the minimum variance of $\hat{a}$ decrease, as shown in Figure 3.10 and Table 3.14. For samples with $r \geq 5$, the optimal fraction for a is nearest to 0.3.

Similarly, nine plots of $Var\left[\hat{b}\right]$ against κ are shown in Figure 3.11; whereas the optimal κ for estimating b and the minimum variance of $\hat{b}$ for each r considered are presented in Table 3.15. For samples with $r = 2$, the lowest variance of $\hat{b}$ is attained when $\kappa = 0.6262$; when r is increased to 5, the κ which minimises $Var\left[\hat{b}\right]$ is 0.3983. The optimal fraction for estimating b is larger than the one for a with a very small margin. Obviously, the best fraction decreases when the separation between the two populations increases. We also note that, for $r \geq 3$, $Var\left[\hat{b}\right]$ is increasing for larger r.

Plots of $Var[\hat{p}]$ against κ are shown in Figure 3.12, representing nine different degrees of separation between the two components. In Table 3.16, we show the smallest variance of $\hat{p}$ and its associated best fraction for each r. Both the optimal κ and $Var[\hat{p}]$ are smaller

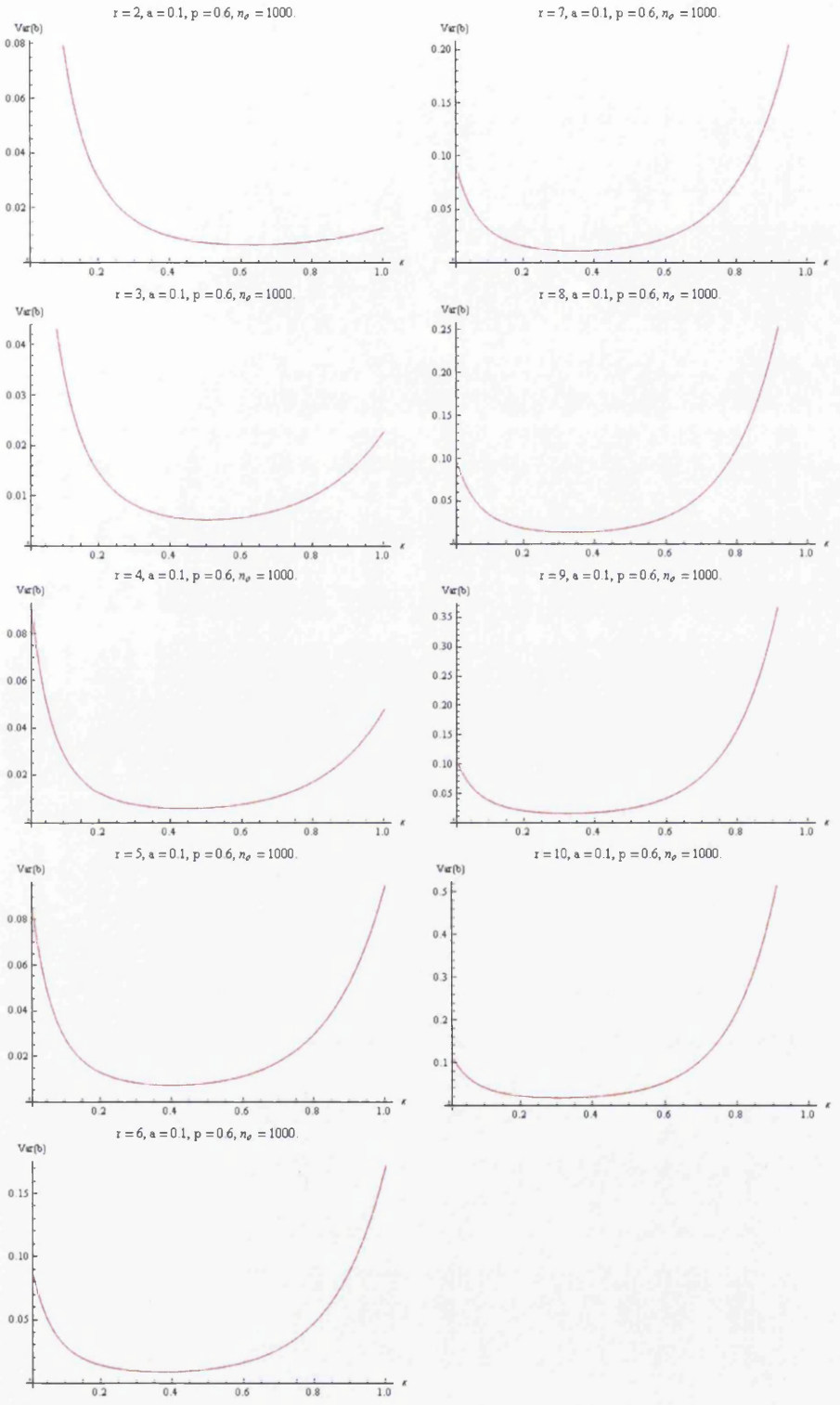


Figure 3.11: Plots of $Var\left[\hat{b}\right]$ versus κ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

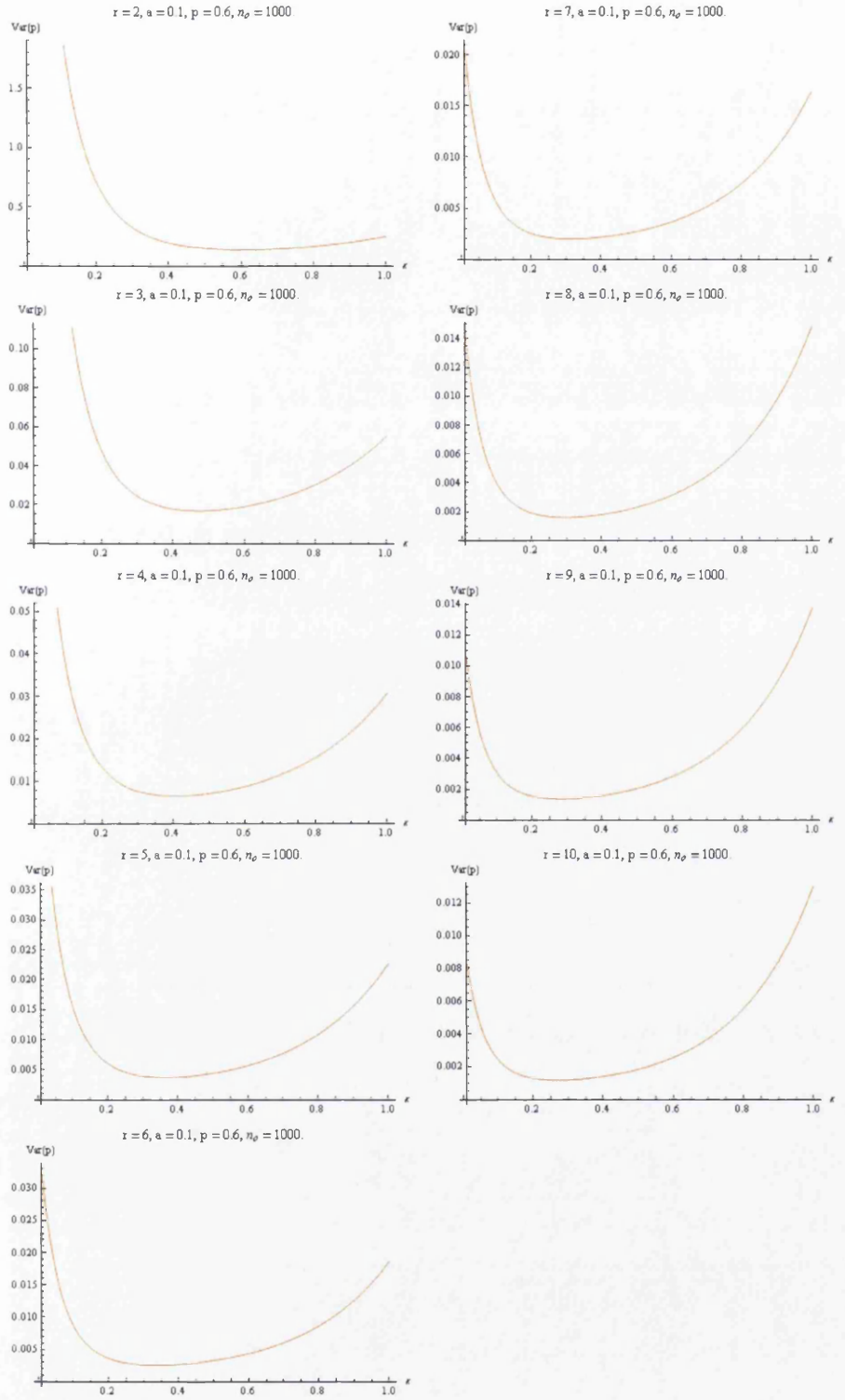


Figure 3.12: Plots of $Var\left[\hat{p}\right]$ versus κ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

r	Theoretical Optimal κ	Practical Optimal κ	Theoretical $Var[\hat{a}]$	Practical $Var[\hat{a}]$
2	0.6	0.9	0.0003	0.0003
3	0.5	0.5	0.0001	0.0001
4	0.4	0.4	6.78×10^{-5}	6.92×10^{-5}
5	0.4	0.4	5.33×10^{-5}	5.47×10^{-5}
6	0.3	0.3	4.44×10^{-5}	4.52×10^{-5}
7	0.3	0.3	3.92×10^{-5}	3.88×10^{-5}
8	0.3	0.3	3.59×10^{-5}	3.63×10^{-5}
9	0.3	0.3	3.37×10^{-5}	3.41×10^{-5}
10	0.3	0.3	3.20×10^{-5}	3.23×10^{-5}

Table 3.17: Theoretical and simulated minimum variance of fractional moment estimator $\hat{a}$ given by the optimal κ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical Optimal κ	Practical Optimal κ	Theoretical $Var\left[\hat{b}\right]$	Practical $Var\left[\hat{b}\right]$
2	0.6	0.6	0.0064	14
3	0.5	0.4	0.0051	0.0079
4	0.4	0.4	0.0060	0.0072
5	0.4	0.4	0.0071	0.0084
6	0.4	0.4	0.0087	0.0099
7	0.3	0.3	0.0108	0.0117
8	0.3	0.3	0.0126	0.0138
9	0.3	0.3	0.0147	0.0156
10	0.3	0.3	0.0169	0.0183

Table 3.18: Theoretical and simulated minimum variance of fractional moment estimator $\hat{b}$ given by the optimal κ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

for larger separation. For $r = 2$, the best κ is 0.6064 and the resulting minimum $Var[\hat{p}]$ is 0.1318; when r is increased to 5, the best κ is 0.3637 and the minimum variance of $\hat{p}$ has a value of 0.0037. The optimal κ's for $Var[\hat{p}]$ have values in between the optimal κ for $Var[\hat{a}]$ and the optimal κ for $Var\left[\hat{b}\right]$.

It is now appropriate to reinforce the theoretical results above with some simulations. For each r, 10000 samples, each of size $n_o = 1000$, were generated from the specified exponential mixture distribution. Every data set was fitted with fractional moment estimators, in which we considered ten values of κ ranging from 0.1 to 1 with an increment of 0.1. From the simulation results, we found the minimum variances of the estimators for a, b and p and recorded the value of κ which gives the minimum variance. The results are presented from Tables 3.17 to 3.19. For simplicity, we round up the theoretical values of κ and the variances

r	Theoretical Optimal κ	Practical Optimal κ	Theoretical $Var[\hat{p}]$	Practical $Var[\hat{p}]$
2	0.6	0.9	0.1318	0.0702
3	0.5	0.5	0.0168	0.0161
4	0.4	0.4	0.0065	0.0065
5	0.4	0.4	0.0037	0.0037
6	0.3	0.3	0.0026	0.0026
7	0.3	0.3	0.0019	0.0019
8	0.3	0.3	0.0016	0.0016
9	0.3	0.3	0.0013	0.0014
10	0.3	0.3	0.0012	0.0012

Table 3.19: Theoretical and simulated minimum variance of fractional moment estimator $\hat{p}$ given by the optimal κ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

of the estimators shown in Tables 3.14 to 3.16. The agreement between observed variances of the derived estimates and theoretical variances of the estimators is very good for all three parameters considered, except when $r = 2$.

The good agreement between the theoretical and practical variance of estimator allows us to conclude the best fraction which yields the minimum variation in the estimators. We have learned that, when the separation between the two components of a mixture is small, we should use a large fraction. For instance, when $r = 2$, we use $\kappa = 0.6$ to ensure good precision of the estimates; for r greater than 6, it is almost certain that the value of $\kappa = 0.3$ will provide us with the best estimates of a, b and p.

In real life, the degree of separation between the two components in a mixture is unknown, how do we know which κ is the best, and which set of parameter estimates have the highest precision? We hereby suggest users a way to confirm the right choice of κ. For illustration, we estimated a single data set, consisting of 1000 observations, simulated from a mixture of two exponential distributions with true parameters $a = 0.1$, $b = 0.5$ and $p = 0.6$ using ten values of κ ranging from 0.1 to 1, with an increment of 0.1. Therefore, we have ten sets of estimates $\hat{\boldsymbol{\Theta}} = \left(\hat{a}, \hat{b}, \hat{p}\right)$, and we substituted each $\hat{\boldsymbol{\Theta}}$ into (3.87) to obtain $Var[\hat{a}]$, $Var\left[\hat{b}\right]$ and $Var[\hat{p}]$ for each set of estimates. In Figure 3.13, we plot these variances along with the asymptotic variances given by the true values, for a, b and p respectively. We can see that both versions of variances have similar patterns, and the optimal κ's (which has a value between 0.3 and 0.4) are similar in both cases. Although the conformity between the theory and practice for b is not as strong as the other two parameters, the excellent agreements for a and p are sufficient for one to discover the best κ for the estimation problem. Therefore, in practice, we should estimate a data set a few times with different values of κ and choose the set of estimates that give the smallest $Var[\hat{a}]$, $Var\left[\hat{b}\right]$ and $Var[\hat{p}]$ in (3.87). By doing this, not only the precision of estimates is guaranteed, at the same time we also

get an intuition about the degree of difference between the populations, judging from the optimal κ: if the best κ appears to be small and the counterpart $\hat{r}$ is large, then we know the estimates are plausible.

3.3.5 Discussion

Like the method of ordinary moments, the fractional moment estimator may return estimates in complex forms. We learned, from our simulation experiments, that the estimates of b can be extremely large/small. Hence, $Var\left[\hat{b}\right]$ is large due to the existence of some outliers (over-estimates of b), especially for samples with small number of observations and small separation between the two components. We now study the reasons we obtain these unreasonable estimates of b.

We know that if s^2 is less than $4t$ in (3.84) and (3.85), then we do not get real roots. Indeed, there is another reason which causes the estimates using fractional moments to be complex numbers: if $\frac{1}{\kappa}$ is not an integer, then we will get complex estimates if x and y are negative. Therefore, when $\kappa = 0.1$, $\kappa = 0.5$ and $\kappa = 1$, we do not get complex estimates even if x and y are negative.

From the simulation results, we can see that estimating b is causing more problems than the other two parameters. The variance of $\hat{b}$ is normally very large due to a few extremely large or extremely small estimates of b. From (3.86), b is given by $y^{-\frac{1}{\kappa}}$. When $\frac{1}{\kappa}$ is an integer (e.g. when κ is 0.1 or 0.5), if y is a very small positive/negative value near the origin, then the estimate of $b$is very large (see Figures 3.14 and 3.15). Conversely, if the absolute value of y is very large, then the estimate of b is very small.

When $\frac{1}{\kappa}$ is not an integer (e.g. when κ is 0.3 or 0.6), we will get complex estimates when x and y are negative. If y is a very small positive value near the origin, then estimates of b will be extremely large (see Figure 3.16). Conversely, b approaches to zero when y increases.

In Section 3.3.2, we have seen that the ordinary moment estimator appears to have much lower variance of $\hat{b}$ compared to the fractional moment estimator when the sample size is small. In Figure 3.17, we see three plots of b against y representing different κ. Compared to ordinary moments, when fractional moments are used, b is more sensitive to y and is more likely to have a large value when $\hat{y}$ is smaller than the true y.

The estimates of s and t can deviate from the true values by a large extent even when the practical values of the fractional moments are close to the true values. When the values of s and t differ greatly from the true values, the estimates of x and y are no longer the ideal values for estimation. This is the reason we get very large/small estimates of b.

Our simulation results show that the fractional moment estimator is actually a good method for large samples, especially when the separation between the two components is large. As seen in Tables 3.11 to 3.13, by replacing ordinary moments with fractional moments, $Var\left[\hat{b}\right]$ is greatly reduced when $n_o = 1000$. We have successfully found the optimal κ for the estimation of each parameter in the previous subsection; the conformity

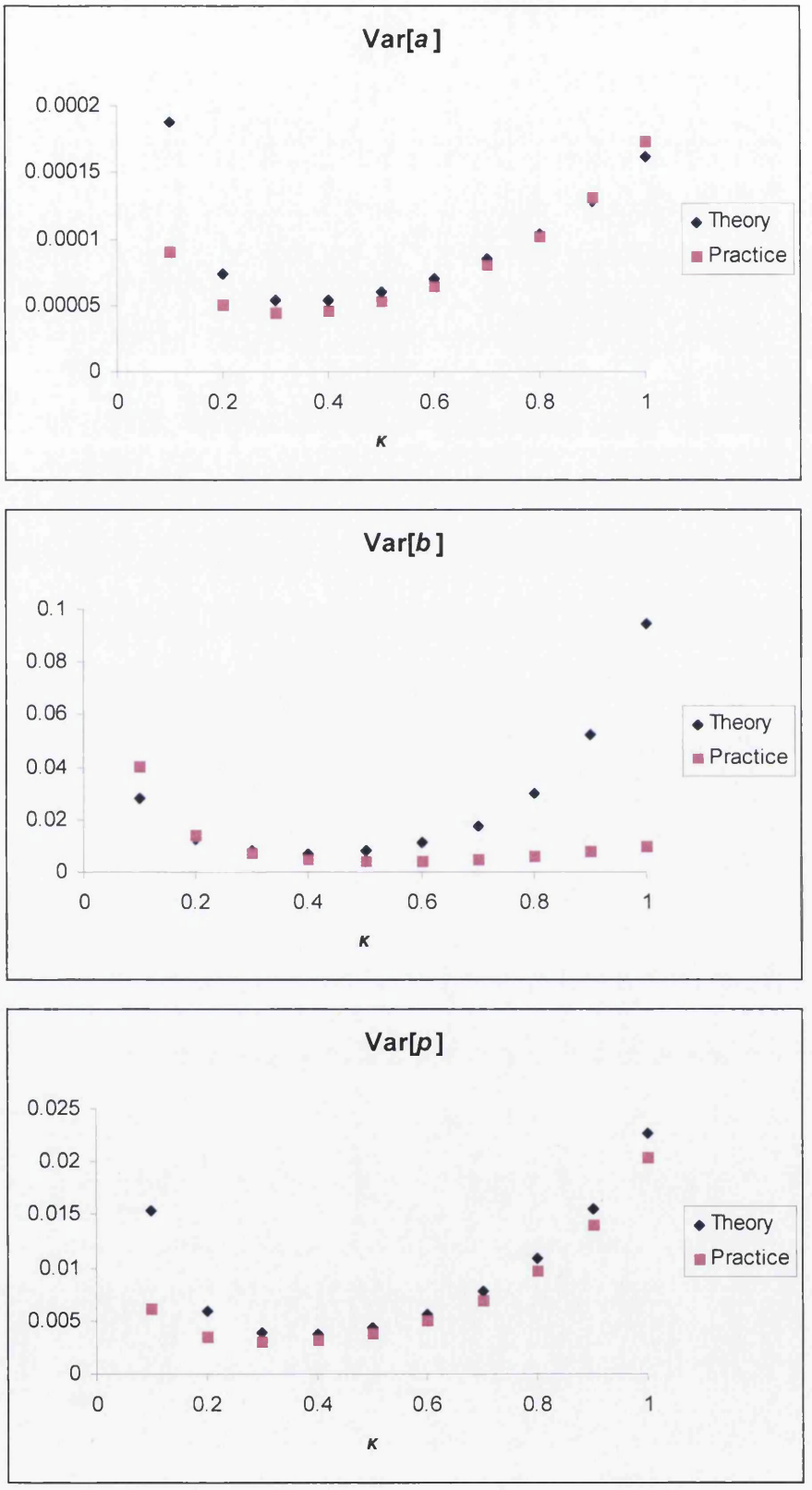


Figure 3.13: Asymptotic variance of the fractional moment estimator given by true parameters and parameter estimates versus κ, based on a data set, consisting of 1000 observations, simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$.

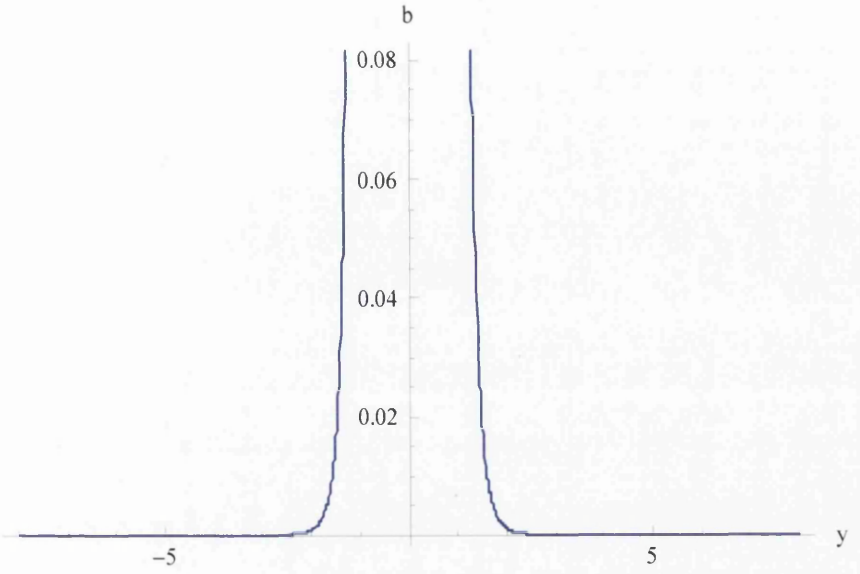


Figure 3.14: Plot of b versus y when $\kappa = 0.1$.

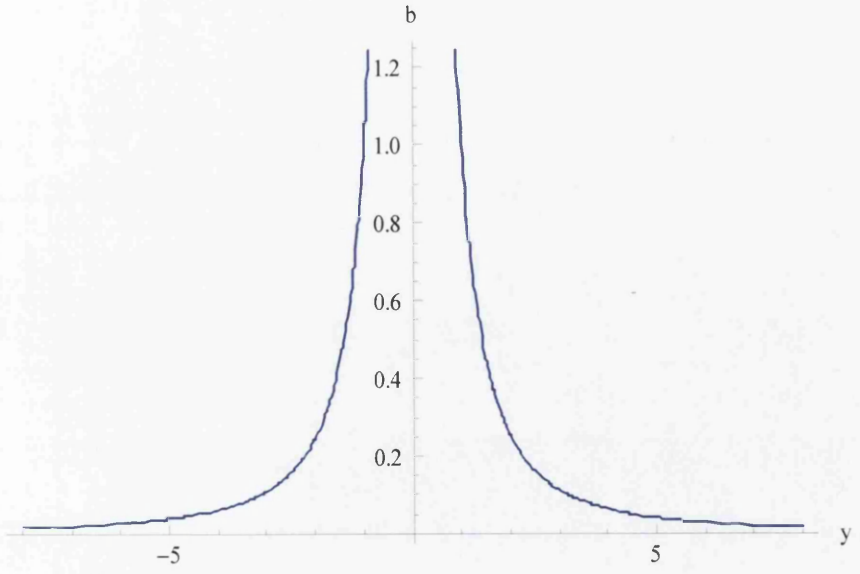


Figure 3.15: Plot of b versus y when $\kappa = 0.5$.

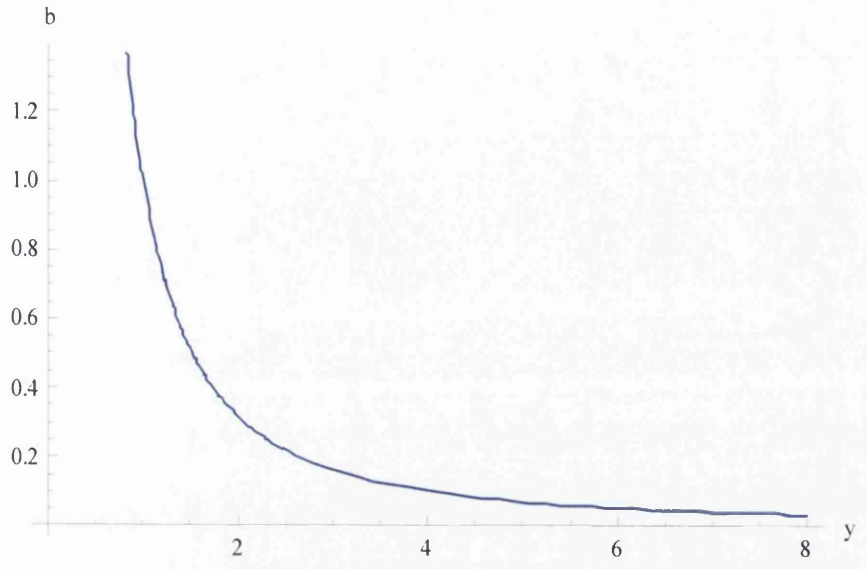


Figure 3.16: Plot of b versus y when $\kappa = 0.6$.

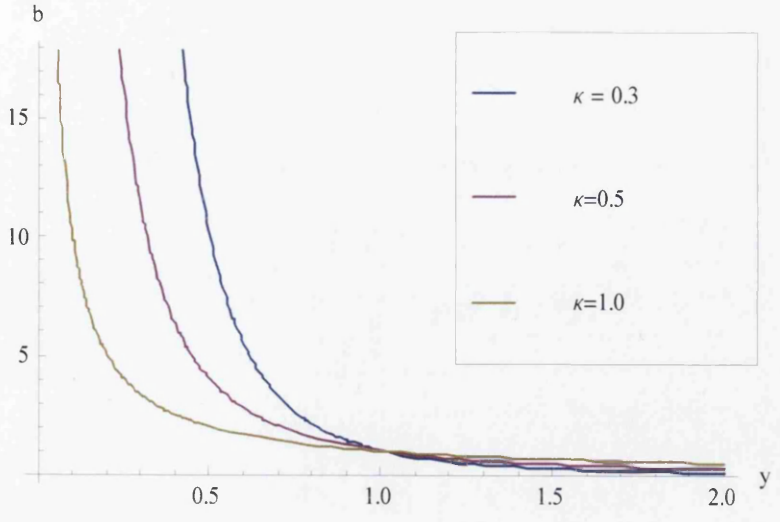


Figure 3.17: Plot of b versus y for various κ

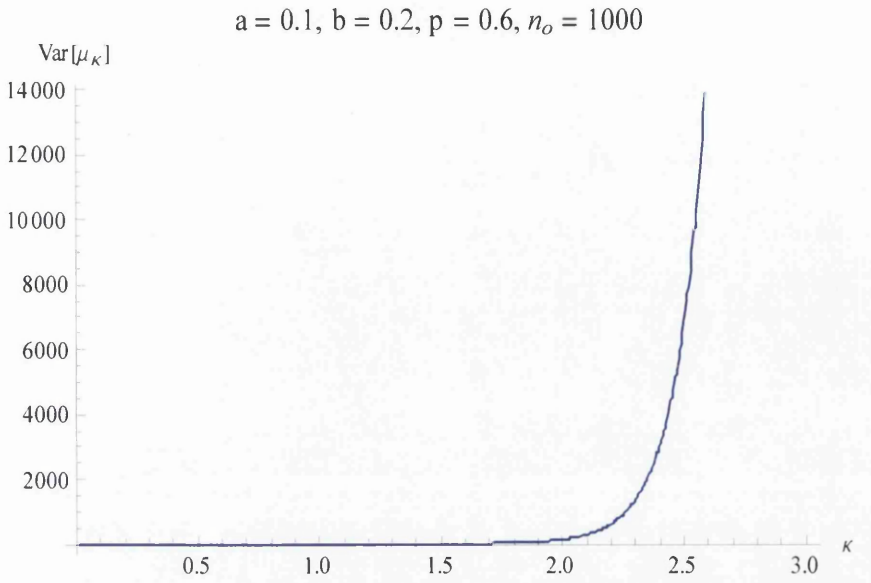


Figure 3.18: Plot of $Var\left[\mu_\kappa\right]$ versus κ for a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$, $p = 0.6$ and $n_o = 1000$.

between the theoretical and the simulated variance of estimator is excellent when the sample size is large. We also found that one of the reasons the ordinary moment estimator is inferior to the fractional moment estimator is $Var\left[\hat{\mu}_\kappa\right]$ increases with κ, as seen in Figure 3.18. Since the method of ordinary moments make use of μ_1, μ_2 and μ_3, and $Var\left[\mu_2\right]$, for instance when $r = 2$, $p = 0.6$ and $n_o = 1000$, is 130 and $Var\left[\mu_3\right]$ is 421290, these large variances of the moments have negative impact on the accuracy of the parameter estimates. On the other hand, for example, when κ is 0.6, $Var\left[\mu_{2\kappa}\right]$ is 0.3233 whereas $Var\left[\mu_{3\kappa}\right]$ is 28, the variances of the fractional moments are significantly smaller than the ones of the ordinary moments. This therefore explains why the fractional moment estimators are better than the traditional moment estimators.

As a conclusion, the method of fractional moments is a promising parameter estimation method for samples with a large number of observations. To obtain the estimates with the highest precision, we suggest the use of the optimal fractions in Tables 3.14, 3.15 and 3.16, if the separation between the two components r is roughly known prior the estimation. We have also suggested a useful way to ensure the precision of the parameter estimates if r is unknown. One can simply estimate a raw sample with ten values of κ ranging from 0.1 to 1, and substitute the resulted ten sets of estimates to (3.87). The set of estimates that gives the smallest $Var\left[\hat{\boldsymbol{\Theta}}\right]$ should be chosen as the final estimates.

For simplicity, we fixed the second fractional moment z_{κ_2} as $z_{2\kappa}$ and the third fractional moment z_{κ_3} as $z_{3\kappa}$ for our study. We speculate that the efficiency of the fractional moment estimator may be increased if the values of these two fractions (κ_2 and κ_3) are allowed to vary.

3.4 The Method of Attenuated Moments

3.4.1 Introduction

The method of attenuated moments, constructed by Jalali (2005c) is a modified version of the method of moments. The theoretical exposition in his subsection follows this paper. With a small attenuation made on the ordinary moments, the ratio between the moments of the two components of the mixture exponential is under control. Let κ and c be positive (real) numbers, a c-attenuated moment of order κ is the expectation

$$\mu_{\kappa}(c) = E\left[T^{\kappa}\exp(-cT)\right], \tag{3.90}$$

where T is an observation from a random sample. It is worth noting that this is a combination of the method of moments with the method of Laplace transforms, where the latter is a generalised method of moments which equates the theoretical and empirical Laplace transform at a set of values of the transform variables:

$$E\left[\exp(-cT)\right] = \frac{1}{n_o}\sum_{i=1}^{n_o}\exp(-ct_i).$$

For more information about the method of Laplace transforms, readers are referred to Yao & Morgan (1999) and Besbeas & Morgan (2004).

In general, a free mixture of m exponential densities with PDF

$$f(t;\boldsymbol{\Theta}) = \sum_{j=1}^{m} p_j\theta_j\exp(-\theta_j t), \tag{3.91}$$

where $f(t) \geq 0$ for all t, and $\sum_{j=1}^{m} p_j = 1$. From (3.90), the c-attenuated moment of order κ of a random variable T which is drawn from the distribution described in (3.91) is defined as

$$\mu_{\kappa}(c) = \Gamma(\kappa+1)\sum_{j=1}^{m}\frac{p_j\theta_j}{(c+\theta_j)^{\kappa+1}}, \tag{3.92}$$

whereas the *normalised* attenuated moment is given by

$$z_{\kappa}(c) = \sum_{j=1}^{m}\frac{p_j\theta_j}{(c+\theta_j)^{\kappa+1}}. \tag{3.93}$$

For estimation purposes, we consider the following system of $2m$ equations

$$z_{\ell\kappa} = \sum_{j=1}^{m}\delta_j x_j^{\ell} \qquad \text{for } \ell = 0, 1, ..., 2m-1, \tag{3.94}$$

where

$$\delta_j = \frac{p_j \theta_j}{c + \theta_j} \tag{3.95}$$

and

$$x_j = (c + \theta_j)^{-\kappa} . \tag{3.96}$$

Note that we first treat δ_j's and x_j's as independent constants.

Lemma 5 *(known) The solutions x_j of the equations in* (3.94) *are the roots of the m^{th} order determinantal equation*

$$\det \begin{bmatrix} z_0 & z_\kappa & z_{2\kappa} & \dots & z_{m\kappa} \\ z_\kappa & z_{2\kappa} & z_{3\kappa} & \dots & z_{(m+1)\kappa} \\ \dots & \dots & \dots & \dots & \dots \\ z_{(m-1)\kappa} & z_{m\kappa} & z_{(m+1)\kappa} & \dots & z_{(2m+1)\kappa} \\ 1 & u & u^2 & \dots & u^m \end{bmatrix} = 0 \tag{3.97}$$

Proof. *Let $\mathbf{V}$ be the Vandermonde matrix based on x_j's:*

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_m \\ x_1^2 & x_2^2 & x_3^2 & \dots & x_m^2 \\ \dots & \dots & \dots & \dots & \dots \\ x_1^{m-1} & x_2^{m-1} & x_3^{m-1} & \dots & x_m^{m-1} \end{bmatrix} . \tag{3.98}$$

It is well known that

$$\det [\boldsymbol{V}] = \prod_{j,k \ s.t. \ k>j} (x_k - x_j) . \tag{3.99}$$

(3.97) *can be expressed as the product of the determinants of the following three matrices*

$$\begin{aligned} \boldsymbol{A} &= \begin{bmatrix} 1 & 1 & \dots & 1 & 0 \\ x_1 & x_2 & \dots & x_m & 0 \\ \dots & \dots & \dots & \dots & \dots \\ x_1^{m-1} & x_2^{m-1} & \dots & x_m^{m-1} & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}, \\ \boldsymbol{B} &= Diag \begin{bmatrix} \delta_1 & \delta_2 & \dots & \delta_m & 1 \end{bmatrix}, \\ \boldsymbol{C} &= \begin{bmatrix} 1 & x_1 & \dots & x_1^{m-1} & x_1^m \\ 1 & x_2 & \dots & x_2^{m-1} & x_2^m \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_m & \dots & x_m^{m-1} & x_m^m \\ 1 & u & \dots & u^{m-1} & u^m \end{bmatrix} . \end{aligned}$$

The determinant of matrix $\mathbf{A}$ is equivalent to the determinant of the Vandermonde matrix

$\boldsymbol{V}$, *given by* (3.99)

$$\det[\boldsymbol{A}] = \prod_{j,k \text{ s.t. } k>j} (x_k - x_j).$$

Matrix $\boldsymbol{B}$ *is a diagonal matrix so its determinant is the product of all its diagonal elements:*

$$\det[\boldsymbol{B}] = \delta_1 \delta_2 ... \delta_m.$$

The determinant of matrix $\boldsymbol{C}$ *is*

$$\det[\boldsymbol{C}] = (u - x_m)(u - x_{m-1}) ... (u - x_1) \prod_{j,k \text{ s.t. } k>j} (x_k - x_j).$$

It follows that

$$\begin{aligned} & \det \begin{bmatrix} z_0 & z_\kappa & z_{2\kappa} & \dots & z_{m\kappa} \\ z_\kappa & z_{2\kappa} & z_{3\kappa} & \dots & z_{(m+1)\kappa} \\ \dots & \dots & \dots & \dots & \dots \\ z_{(m-1)\kappa} & z_{m\kappa} & z_{(m+1)\kappa} & \dots & z_{(2m+1)\kappa} \\ 1 & u & u^2 & \dots & u^m \end{bmatrix} \\ = \; & \det[\boldsymbol{A}] \det[\boldsymbol{B}] \det[\boldsymbol{C}] \\ = \; & \left[\prod_{j,k \text{ s.t. } k>j} (x_k - x_j)^2 \right] [\delta_1 \delta_2 ... \delta_m] [(u - x_m)(u - x_{m-1}) ... (u - x_1)]. \end{aligned}$$

Provided that $x_k \neq x_j$ *and* $\delta_j \neq 0$, (3.97) *is equivalent to* 0 *when and only when* $u = x_1$, $u = x_2$, ..., *or* $u = x_m$. ■

After finding the roots of (3.97), we shall find δ_j's from the first m equations as follows:

$$\begin{bmatrix} \delta_1 & \delta_2 & \delta_3 & \dots & \delta_m \end{bmatrix}^T = \boldsymbol{V}^{-1} \begin{bmatrix} z_0 & z_k & z_{2k} & \dots & z_{(m-1)k} \end{bmatrix}^T. \tag{3.100}$$

where $\boldsymbol{V}$ is the Vandermonde matrix (3.99). Having found x_j's and δ_j's, we can now find θ_j and p_j :

$$\theta_j = x_j^{-\frac{1}{\kappa}} - c = \frac{1 - c x_j^{\frac{1}{\kappa}}}{x_j^{\frac{1}{\kappa}}} \tag{3.101}$$

and

$$p_j = \frac{\delta_j}{1 - c x_j^{\frac{1}{\kappa}}}. \tag{3.102}$$

Obviously, the weights should sum up to one, hence

$$\sum_{j=1}^{m} \frac{\delta_j}{1 - c x_j^{\frac{1}{\kappa}}} = 1. \tag{3.103}$$

Clearly in the preceding, we have actual parameters and theoretical attenuated moments, and with such parameters, the preceding equations hold precise and unambiguously. It can also be seen from these equations that unless two or more θ_js are equal, V can be inverted and the procedure is continuous. For "large" samples, these sample moments come very "close" to the actual moments, and because of the continuity of our process, the derived (estimated) parameters will also be very "close" to the actual parameters, no matter which equations we use, provided that there are sufficiently many.

Now let $t_1, ..., t_{n_o}$ be a sample of size n_o of our sojourn times. Then the sample c-attenuated fractional moment of order κ is

$$\hat{z}_\kappa(c) = \frac{1}{n_o \Gamma(\kappa+1)} \sum_{i=1}^{n_o} t_i^\kappa \exp(-ct_i). \tag{3.104}$$

When there is no room for confusion, we drop the argument c. Now, if in the above procedures we replace $z_{\ell\kappa}$ by $\hat{z}_{\ell\kappa}$, we obtain, progressively from (3.97) to (3.101), $\hat{x}_j$, $\hat{\delta}_j$, $\hat{\theta}_j$. Following (3.102), if we set

$$\hat{p}^* = \sum_{j=1}^{m} \frac{\hat{\delta}_j}{1 - c\hat{x}_j^{\frac{1}{\kappa}}}, \tag{3.105}$$

then the observed value of this statistic should be close to 1. If it deviates "too much" from 1, the data does not conform to a free mixture of exponentials. If it is close to 1, we calculate the estimates of the weights as follows:

$$\hat{p}_j = \frac{\hat{\delta}_j}{\hat{p}^*(1 - c\hat{x}_j^{\frac{1}{\kappa}})}. \tag{3.106}$$

This concludes the process of estimation based on attenuated fractional moments.

It is worth noting that we first pretend as that the weights are unconstrained, and then we go on to solve $2m$ equations with $2m$ unknowns. Because of the continuity of the procedure we just mentioned, if the sample is infinite, we expect to get exact weights which add up to 1; for large samples, we expect the sum of estimated weights which is "close" to 1. At this point, we have two routes to proceed further:

Route 1: A non elegant way in which we keep everything we found including the sum of weights which is not equal to 1.

Route 2: In which we have chosen to normalise the weights by summing them up and divide each by the sum in (3.105). Of course with these new weights, the sample moment is not equal to the theoretical moment derived from estimates because of these normalisation, but they should be very "close" to each other. We have chosen this route which is a more elegant way to estimate the parameters.

Mixture of Two Exponential Distributions

We now consider the attenuated moment estimation of the parameters of a two-component exponential mixture model with parameter vector $\Theta = (a, b, p)$ and PDF in (3.4). Following the procedures described above, we first set

$$\begin{aligned} \delta_1 &= \frac{pa}{c+a}, \\ \delta_2 &= \frac{(1-p)\,b}{c+b}, \\ x_1 &= (c+a)^{-\kappa}, \\ x_2 &= (c+b)^{-\kappa}. \end{aligned} \tag{3.107}$$

From (3.94), we need a system of four attenuated fractional moments to solve the problem of a two-component mixture exponential density:

$$\begin{aligned} z_0 &= \frac{pa}{c+a} + \frac{(1-p)\,b}{c+b}, \\ z_\kappa &= \frac{pa}{(c+a)^{\kappa+1}} + \frac{(1-p)\,b}{(c+b)^{\kappa+1}}, \\ z_{2\kappa} &= \frac{pa}{(c+a)^{2\kappa+1}} + \frac{(1-p)\,b}{(c+b)^{2\kappa+1}}, \\ z_{3\kappa} &= \frac{pa}{(c+a)^{3\kappa+1}} + \frac{(1-p)\,b}{(c+b)^{3\kappa+1}}. \end{aligned} \tag{3.108}$$

To obtain x_j, we solve the following equation:

$$\det \begin{bmatrix} z_0 & z_\kappa & z_{2\kappa} \\ z_\kappa & z_{2\kappa} & z_{3\kappa} \\ 1 & u & u^2 \end{bmatrix} = 0$$

$$\Leftrightarrow (z_0 z_{2\kappa} - z_\kappa^2)u^2 + (z_\kappa z_{2\kappa} - z_0 z_{3\kappa})u + (z_\kappa z_{3\kappa} - z_{2\kappa}^2) = 0$$

$$\Leftrightarrow u^2 - \frac{z_0 z_{3\kappa} - z_\kappa z_{2\kappa}}{z_0 z_{2\kappa} - z_\kappa^2} u + \frac{z_\kappa z_{3\kappa} - z_{2\kappa}^2}{z_0 z_{2\kappa} - z_\kappa^2} = 0. \tag{3.109}$$

Now, if we let

$$\begin{aligned} s &= \frac{z_0 z_{3\kappa} - z_\kappa z_{2\kappa}}{z_0 z_{2\kappa} - z_\kappa^2}, \\ t &= \frac{z_\kappa z_{3\kappa} - z_{2\kappa}^2}{z_0 z_{2\kappa} - z_\kappa^2}, \end{aligned} \tag{3.110}$$

then (3.109) is in the form of

$$u^2 - su + t = 0. \tag{3.111}$$

Therefore, we can get x_j by finding the roots of the 2^{nd} order determinantal equation in (3.111):

$$\begin{aligned} x_1 &= \frac{s+\sqrt{s^2-4t}}{2}, \\ x_2 &= \frac{s-\sqrt{s^2-4t}}{2}. \end{aligned} \tag{3.112}$$

We need to know δ_j in order to estimate the weights. From (3.98), the Vandermonde matrix, V for $m = 2$ is

$$V = \begin{bmatrix} 1 & 1 \\ x_1 & x_2 \end{bmatrix}. \tag{3.113}$$

It follows that

$$\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = V^{-1} \begin{bmatrix} z_0 \\ z_\kappa \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \frac{1}{x_2 - x_1} \begin{bmatrix} x_2 & -1 \\ -x_1 & 1 \end{bmatrix} \begin{bmatrix} z_0 \\ z_\kappa \end{bmatrix} \tag{3.114}$$

$$\Leftrightarrow \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \frac{1}{x_2 - x_1} \begin{bmatrix} z_0 x_2 - z_\kappa \\ z_\kappa - z_0 x_1 \end{bmatrix}.$$

Since we have x_j's and δ_j's, we can now find the estimates of θ_j's and p_j's:

$$\begin{aligned} a &= \frac{1}{x_1^{\frac{1}{\kappa}}} - c, \\ b &= \frac{1}{x_2^{\frac{1}{\kappa}}} - c, \\ p^* &= \frac{\delta_1}{(1 - c x_1^{\frac{1}{\kappa}})} + \frac{\delta_2}{(1 - c x_2^{\frac{1}{\kappa}})}, \\ p &= \frac{\delta_1}{p^*(1 - c x_1^{\frac{1}{\kappa}})}, \text{ if } p^* \text{ is close to 1.} \end{aligned} \tag{3.115}$$

By replacing z_κ by the sample moments $\hat{z}_\kappa$ (as in (3.104)), we obtain the attenuated moment estimators from (3.115) for a sample drawn from a mixture of two exponential distributions.

In Table 3.20, we illustrate how the variation between the moments is further controlled when an attenuation is applied to a fractional moment. Like before, the first component has a rate parameter $a = 0.1$, the second component's rate parameter is $b = 1$ and the mixing probability of the first component is 0.6. Recall from Table 3.9 that, when the method of ordinary moments is used, the ratio of the third moment of the first exponent to the second exponent is 1000. This causes poor estimation of the second rate parameter because the moment of the second component is too small and hence it pales into insignificance. As

(κ, c)	$z_{a\kappa}(c)$	$z_{b\kappa}(c)$	$z_{\kappa}(c)$	$\frac{z_{a\kappa}(c)}{z_{b\kappa}(c)}$
$(0.3, 0.03)$	1.4186	0.9623	1.2361	1.4742
$(0.6, 0.03)$	2.6163	0.9538	1.9513	2.7430
$(0.9, 0.03)$	4.8251	0.9454	3.2732	5.1039

Table 3.20: Theoretical moments $z_{\kappa}(c)$ of a sample arising from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$, and the ratio of the moments of the two exponential components.

mentioned in previous section, the fractional moments controls the variation between the moments; the ratio reduced to 10 when $\kappa = \frac{1}{3}$, as shown in Table 3.10. In our example here, we make an attenuation $c = 0.03$ to the moments and the ratio the third moments $\frac{z_{a\kappa}(c)}{z_{b\kappa}(c)}$ is further reduced to 5.

3.4.2 Simulation Results

In this subsection, we look at the performance of the method of attenuated moments for the estimation problem of mixtures of two exponential distributions. Like before, we consider three sets of parameters $\Theta = (0.1, 0.1r, 0.6)$, where $r = (2, 5, 10)$ representing three different degrees of separation between the two exponential components. For each case, we investigate the estimator's performance on different sample sizes, $n_o = (10, 15, 20, 50, 1000)$. Our main interest is the optimal combination of fraction κ and attenuation c. We consider ten values of $\kappa = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$; for each κ, we consider ten values of attenuation, $c = (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1)$. For every set of parameters, we simulate 10000 independent samples each consisting of n_o observations and the "unknown" parameter Θ is estimated for each data set based on the method of attenuated moments with the 100 different combinations of κ and c. The minimum measures of errors are presented in Tables 3.21 to 3.23. From these three tables, we observe that, for all r considered, $Var[\hat{a}]$, $Var\left[\hat{b}\right]$ and $Var[\hat{p}]$ decrease as the sample size increases. $Var\left[\hat{b}\right]$ are significantly large for small sample size. Only when the separation between the two components is large ($r = 10$), $Var\left[\hat{b}\right]$ becomes under control when $n_o = 50$.

For $r = 2$ (see Table 3.21), remarkably, the best combination of κ and c is $(0.9, 0.01)$, in terms of both the bias and the mean square error, for all sample sizes considered by us. Our simulated results suggest that, to estimate b in this case, we should use $\kappa \geq 0.9$ and $c \leq 0.04$; whereas for p, we should use $\kappa \geq 0.9$ and $c \leq 0.02$ for all sample sizes. The pattern of the combination is clear: when the separation between the two components is small ($r = 2$), the best combination has a large fraction (κ is either 0.9 or 1), and a small attenuation (c is between 0.01 and 0.04), except when $n_o = 1000$. For such a small r, the estimation of b is not satisfactory: $Var\left[\hat{b}\right]$ is large, especially for small samples.

When $r = 5$, we can see from Table 3.22 that the best combinations for all three parameters are made up of a large fraction and a small attenuation: for a, we should use

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	2.57×10^{-9} $(0.3, 0.01)$	4.20×10^{-7} $(0.5, 0.04)$	1.80×10^{-7} $(0.5, 0.04)$	4.78×10^{-8} $(0.2, 0.04)$	2.42×10^{-6} $(1, 0.01)$
$(\bar{\hat{b}} - b)^2$	1.36×10^{-5} $(1, 0.01)$	6.71×10^{-8} $(1, 0.09)$	0.0020 $(1, 0.01)$	0.0032 $(1, 0.01)$	9.95×10^{-6} $(1, 0.04)$
$(\bar{\hat{p}} - p)^2$	1.25×10^{-7} $(0.8, 0.07)$	4.64×10^{-6} $(0.6, 0.02)$	5.96×10^{-8} $(0.6, 0.03)$	1.55×10^{-6} $(0.6, 0.05)$	4.63×10^{-6} $(0.4, 0.09)$
$Var[\hat{a}]$	0.0017 $(0.9, 0.01)$	0.0017 $(0.9, 0.01)$	0.0015 $(0.9, 0.01)$	0.0013 $(0.9, 0.01)$	0.0003 $(0.9, 0.01)$
$Var[\hat{b}]$	322 $(1, 0.01)$	96 $(1, 0.04)$	50 $(0.9, 0.02)$	65 $(0.9, 0.02)$	2 $(0.9, 0.07)$
$Var[\hat{p}]$	0.2919 $(0.9, 0.01)$	0.2585 $(1, 0.02)$	0.2729 $(1, 0.02)$	0.1793 $(0.1, 0.02)$	0.0697 $(0.9, 0.01)$
$MSE[\hat{a}]$	0.0018 $(0.9, 0.01)$	0.0018 $(0.9, 0.01)$	0.0016 $(0.9, 0.01)$	0.0015 $(0.9, 0.01)$	0.0003 $(0.9, 0.01)$
$MSE[\hat{b}]$	322 $(1, 0.01)$	96 $(1, 0.04)$	51 $(0.9, 0.02)$	65 $(0.9, 0.02)$	2 $(0.9, 0.07)$
$MSE[\hat{p}]$	0.2938 $(0.9, 0.01)$	0.2609 $(1, 0.02)$	0.2746 $(1, 0.02)$	0.1918 $(0.9, 0.02)$	0.0703 $(0.9, 0.01)$

Table 3.21: Performance of the method of attenuated moments for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	1.56×10^{-7} $(0.6, 0.03)$	2.34×10^{-8} $(0.3, 0.01)$	3.84×10^{-9} $(1, 0.07)$	9.70×10^{-10} $(1, 0.06)$	8.52×10^{-11} $(0.9, 0.07)$
$(\bar{\hat{b}} - b)^2$	0.0013 $(1, 0.1)$	0.0680 $(1, 0.05)$	0.0045 $(1, 0.04)$	0.0006 $(1, 0.04)$	4.00×10^{-5} $(0.6, 0.08)$
$(\bar{\hat{p}} - p)^2$	2.26×10^{-7} $(0.6, 0.1)$	7.99×10^{-7} $(0.3, 0.08)$	2.71×10^{-8} $(0.3, 0.07)$	4.74×10^{-7} $(0.1, 0.08)$	2.54×10^{-9} $(0.8, 0.05)$
$Var[\hat{a}]$	0.0065 $(0.9, 0.04)$	0.0038 $(0.9, 0.05)$	0.0026 $(0.9, 0.01)$	0.0010 $(0.8, 0.01)$	4.57×10^{-5} $(0.6, 0.03)$
$Var[\hat{b}]$	401 $(1, 0.06)$	429 $(1, 0.05)$	312 $(1, 0.01)$	203 $(0.7, 0.07)$	0.0058 $(0.7, 0.06)$
$Var[\hat{p}]$	0.1790 $(1, 0.01)$	0.1497 $(1, 0.03)$	0.1157 $(0.9, 0.04)$	0.0506 $(0.8, 0.01)$	0.0030 $(0.7, 0.06)$
$MSE[\hat{a}]$	0.0066 $(0.9, 0.04)$	0.0039 $(0.8, 0.02)$	0.0027 $(0.9, 0.01)$	0.0010 $(0.8, 0.01)$	4.57×10^{-5} $(0.6, 0.03)$
$MSE[\hat{b}]$	401 $(1, 0.06)$	429 $(1, 0.05)$	312 $(1, 0.01)$	204 $(0.7, 0.07)$	0.0059 $(0.7, 0.06)$
$MSE[\hat{p}]$	0.1912 $(1, 0.01)$	0.1547 $(1, 0.03)$	0.1230 $(1, 0.03)$	0.0514 $(0.8, 0.01)$	0.0030 $(0.7, 0.06)$

Table 3.22: Performance of the method of attenuated moments for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{\bar{a}} - a)^2$	1.85×10^{-7} $(0.9, 0.03)$	1.53×10^{-7} $(0.8, 0.03)$	8.37×10^{-11} $(0.7, 0.05)$	1.27×10^{-10} $(0.5, 0.05)$	2.49×10^{-11} $(0.6, 0.07)$
$(\hat{\bar{b}} - b)^2$	0.1382 $(1, 0.03)$	0.1323 $(1, 0.01)$	0.2633 $(1, 0.1)$	0.0053 $(1, 0.08)$	2.41×10^{-5} $(0.5, 0.1)$
$(\hat{\bar{p}} - p)^2$	1.12×10^{-6} $(0.3, 0.03)$	2.02×10^{-7} $(0.5, 0.06)$	1.30×10^{-5} $(0.5, 0.1)$	6.10×10^{-8} $(0.7, 0.05)$	1.31×10^{-10} $(0.5, 0.05)$
$Var[\hat{a}]$	0.0025 $(0.9, 0.01)$	0.0014 $(0.9, 0.01)$	0.0011 $(0.9, 0.01)$	0.0005 $(0.9, 0.01)$	2.84×10^{-5} $(0.4, 0.02)$
$Var[\hat{b}]$	454 $(0.9, 0.03)$	155 $(1, 0.01)$	204 $(0.9, 0.01)$	1.5620 $(0.3, 0.03)$	0.0122 $(0.6, 0.09)$
$Var[\hat{p}]$	0.0554 $(0.8, 0.01)$	0.0407 $(0.8, 0.03)$	0.0279 $(0.9, 0.02)$	0.0159 $(0.9, 0.01)$	0.0009 $(0.7, 0.1)$
$MSE[\hat{a}]$	0.0025 $(0.9, 0.01)$	0.0014 $(0.9, 0.01)$	0.0011 $(0.9, 0.01)$	0.0005 $(0.9, 0.01)$	2.84×10^{-5} $(0.4, 0.02)$
$MSE[\hat{b}]$	457 $(0.9, 0.03)$	156 $(1, 0.01)$	207 $(0.9, 0.01)$	1.6436 $(0.3, 0.03)$	0.0123 $(0.6, 0.09)$
$MSE[\hat{p}]$	0.0581 $(0.8, 0.01)$	0.0422 $(0.7, 0.01)$	0.0299 $(0.9, 0.02)$	0.0165 $(0.9, 0.01)$	0.0009 $(0.7, 0.1)$

Table 3.23: Performance of the method of attenuated moments for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$ for different sample size n_o.

$0.6 \leq \kappa \leq 0.9$ and $0.01 \leq c \leq 0.05$, for b, $0.7 \leq \kappa \leq 1$ and $0.05 \leq c \leq 0.07$ whereas for p, $0.7 \leq \kappa \leq 1$ and $0.01 \leq c \leq 0.06$.

When $r = 10$, as seen in Table 3.23, the best combination of κ and c for estimating a, in terms of both the variance and the mean square error, is $(0.9, 0.01)$ for all samples with size $n_o \leq 50$; whereas if we use $\kappa = 0.4$ and $c = 0.02$ to estimate a from a sample of size $n_o = 1000$, $Var[\hat{a}] = 2.84 \times 10^{-5}$ is the minimum variance we can obtain. For b, when $n_o \leq 20$, the best κ is either 0.9 or 1, with $0.01 \leq c \leq 0.03$; the best fraction decreases when the sample size increases: when $n_o = 1000$, the ideal combination for b is $\kappa = 0.6$ and $c = 0.09$. For p, the optimal combination is a large fraction $(0.8 \leq \kappa \leq 0.9)$ and a small attenuation $(0.01 \leq c \leq 0.03)$ for $n_o \leq 50$; whereas $\kappa = 0.7$ and $c = 0.1$ is the best combination for estimating p from large data set with $n_o = 1000$.

Overall speaking, we are satisfied with the performance of the attenuated moment estimator. Encouragingly, its bias^2 of $\hat{a}$ and $\hat{p}$ are very close to zero even when the data set has a small number of observations and a small r. Although the variances of $\hat{b}$ are quite large when n_o is small, as expected, we are happy to see that the estimation of b is nicely improved for samples with a large size and a large separation between the two distributions. It is also worth noting that the variances of moment estimators are greatly reduced by using attenuated moments. As an example, for data sets with $r = 5$ and $n_o = 1000$, $Var[\hat{b}]$ was 3.74×10^5 when ordinary moments were employed $(\kappa = 1)$ and 0.0079 when fractional

moments were used ($\kappa = 0.4$); however, when attenuated moments were used ($\kappa = 0.7$ and $c = 0.06$), the variance of $\hat{b}$ is reduced to 0.0058; not to mention that the bias of attenuated moment estimators are also significantly smaller than the ones given by the ordinary moment estimator and the fractional moment estimator.

3.4.3 Asymptotic Covariance Matrix of the Attenuated Fractional Moment Estimator

In this subsection, we calculate the asymptotic covariance matrix of the attenuated moment estimator by following the procedure in Section 1.5.5. As we mentioned before, it is impossible to find an explicit form of such a matrix. Therefore, we use Taylor's Expansion to find an approximation to the theoretical matrix.

The attenuated fractional moments of a mixture of two exponential distributions are in the form of

$$\mu_{\kappa}(c) = E[T^{\kappa}\exp(-cT)] = \Gamma(\kappa+1)\left[\frac{pa}{(c+a)^{\kappa+1}} + \frac{(1-p)b}{(c+b)^{\kappa+1}}\right]. \tag{3.116}$$

The Jacobian matrix $\boldsymbol{D}[\boldsymbol{\Theta}]$ has entries $d_{ij} = \dfrac{\partial \mu_{\kappa_i}(c)}{\partial \Theta_j}$, where

$$\begin{aligned}
\frac{\partial \mu_{\kappa_i}(c)}{\partial a} &= p\Gamma(\kappa_i+1)\frac{c-a\kappa_i}{(c+a)^{\kappa_i+2}}, \\
\frac{\partial \mu_{\kappa_i}(c)}{\partial b} &= (1-p)\Gamma(\kappa_i+1)\frac{c-b\kappa_i}{(c+b)^{\kappa_i+2}}, \\
\frac{\partial \mu_{\kappa_i}(c)}{\partial p} &= \Gamma(\kappa_i+1)\left[\frac{a}{(c+a)^{\kappa_i+1}} - \frac{b}{(c+b)^{\kappa_i+1}}\right],
\end{aligned} \tag{3.117}$$

for $i = 1, 2, 3$ (for simplicity, we set κ_2 as 2κ and κ_3 and 3κ for our investigation), while the covariance matrix of the attenuated fractional moments $\boldsymbol{V}[\hat{\boldsymbol{\mu}}]$ has entries

$$\begin{aligned}
& Cov\left[T^{\kappa}\exp(-cT), T^{\ell}\exp(-cT)\right] \\
&= \frac{1}{n_o}\left[E\left[T^{\kappa}\exp(-cT)\,T^{\ell}\exp(-cT)\right] - E\left[T^{\kappa}\exp(-cT)\right]E\left[T^{\ell}\exp(-cT)\right]\right] \\
&= \frac{1}{n_o}\left[E\left[T^{\kappa+\ell}\exp(-2cT)\right] - E\left[T^{\kappa}\exp(-cT)\,.T^{\ell}\exp(-cT)\right]\right] \\
&= \frac{1}{n_o}\begin{bmatrix} \Gamma(\kappa+\ell+1)\left[\dfrac{pa}{(2c+a)^{\kappa+\ell+1}} + \dfrac{(1-p)b}{(2c+b)^{\kappa+\ell+1}}\right] \\ -\Gamma(\kappa+1)\Gamma(\ell+1)\left[\dfrac{pa}{(c+a)^{\kappa+1}} + \dfrac{(1-p)b}{(c+b)^{\kappa+1}}\right]\left[\dfrac{pa}{(c+a)^{\ell+1}} + \dfrac{(1-p)b}{(c+b)^{\ell+1}}\right] \end{bmatrix}.
\end{aligned} \tag{3.118}$$

Thus, following the general formula in (1.47), the covariance matrix of the attenuated frac-

tional moment estimator $\boldsymbol{\Theta}$ can be calculated by

$$\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right] \approx \boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1} \boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right] \left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{T}\right)^{-1}, \tag{3.119}$$

where $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$ and $\boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right]$ are obtained from (3.117) and (3.118).

The method of attenuated moments uses four moments to estimate three parameters $\boldsymbol{\Theta} = (a, b, p)$, this means that the Jacobian matrix $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$ is a 4×3 matrix. Since $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$ is not a square matrix, we cannot find its inverse as required in (3.119). In order to solve this, we have attempted two approaches.

First, we assume that there are four parameters to be estimated, by including q to $\boldsymbol{\Theta}$. We amend $\dfrac{\partial \mu_{\kappa_i}(c)}{\partial p}$ and it now is in the form of

$$\frac{\partial \mu_{\kappa_i}(c)}{\partial p} = \Gamma(\kappa_i + 1)\frac{a}{(c+a)^{\kappa_i+1}}, \tag{3.120}$$

$\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$ has now an extra column given by

$$\frac{\partial \mu_{\kappa_i}(c)}{\partial q} = \Gamma(\kappa_i + 1)\frac{b}{(c+b)^{\kappa_i+1}}. \tag{3.121}$$

Hence, the size of every matrix is now 4×4 and we manage to solve (3.119). However, there is a drawback of this approach because we do not really need to estimate q in practice. This might not be the best way to obtain $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ but fortunately the agreements between the theoretical and practical variances are good.

Another way to solve (3.119) is by finding the generalised inverse of $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$, which is given by

$$\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1} = \left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{T} \boldsymbol{D}\left[\boldsymbol{\Theta}\right]\right)^{-1} \boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{T} \tag{3.122}$$

and

$$\left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{T}\right)^{-1} = \boldsymbol{D}\left[\boldsymbol{\Theta}\right] \left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{T} \boldsymbol{D}\left[\boldsymbol{\Theta}\right]\right)^{-1}. \tag{3.123}$$

Having done this, $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1}$ becomes a 3×4 matrix; $\left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{T}\right)^{-1}$ becomes a 4×3 matrix; and hence $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ remains as a 3×3 matrix. Note that (3.122) and (3.123) can be tricky sometimes so we cannot be too optimistic on the approximation with this approach.

We calculate these two versions' theoretical variances of the estimators and compare them to see if they agree with each other. Then, we move on to find conformity between simulated and theoretical results.

3.4.4 Optimal Combination of κ and c

As discussed in the previous subsection, we have two versions of $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ for the method of attenuated moments. In this subsection, we investigate which version has a closer agreement

r	Theoretical $Var[\hat{a}]_Q$	Theoretical $Var[\hat{a}]_{GI}$	Practical $Var[\hat{a}]$
2	0.0003 ($\kappa = 0.97, c = 0.027$)	0.0003 ($\kappa = 0.98, c = 0.028$)	0.0003 ($\kappa = 0.90, c = 0.010$)
5	4.61×10^{-5} ($\kappa = 0.61, c = 0.030$)	4.61×10^{-5} ($\kappa = 0.61, c = 0.030$)	4.57×10^{-5} ($\kappa = 0.60, c = 0.030$)
10	2.90×10^{-5} ($\kappa = 0.42, c = 0.026$)	2.90×10^{-5} ($\kappa = 0.42, c = 0.025$)	2.84×10^{-5} ($\kappa = 0.40, c = 0.020$)

Table 3.24: Theoretical and simulated minimum variance of attenuated moment estimator $\hat{a}$ given by the optimal combination of κ and c for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical $Var\left[\hat{b}\right]_Q$	Theoretical $Var\left[\hat{b}\right]_{GI}$	Practical $Var\left[\hat{b}\right]$
2	0.0056 ($\kappa = 1.02, c = 0.031$)	0.0056 ($\kappa = 1.02, c = 0.031$)	2 ($\kappa = 0.90, c = 0.070$)
5	0.0052 ($\kappa = 0.86, c = 0.079$)	0.0050 ($\kappa = 0.60, c = 0.147$)	0.0058 ($\kappa = 0.70, c = 0.06$)
10	0.0113 ($\kappa = 0.83, c = 0.156$)	0.0106 ($\kappa = 0.69, c = 0.33$)	0.0122 ($\kappa = 0.60, c = 0.09$)

Table 3.25: Theoretical and simulated minimum variance of attenuated moment estimator $\hat{b}$ given by the optimal combination of κ and c for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

to the practical results. From now on, we denote $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]_Q$ as the variance of attenuated moment estimator where we include the parameter $q\,(=1-p)$ to the parameter vector $\boldsymbol{\Theta}$, and $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]_{GI}$ as the variance of attenuated moment estimator where we use the generalised inverse to solve (3.119). It is also our objective to find the best combination of fraction κ and attenuation c that yields the lowest variance of estimator. In Mathematica, we used the function "FindMinimum" to obtain the values of κ and c which minimise the variance of estimator in (3.119).

In order to confirm if the practical results agree with the theoretical results, we make use of the practical simulation results ($n_o = 1000$) from Tables 3.21 to 3.23 to compare the variances of the estimators with the theoretical variances. In Table 3.24, we present both versions' minimum theoretical variance of estimator a and the theoretical best combination of fraction κ and attenuation c for $r = 2$, 5 and 10 when $n_o = 1000$. We also show the practical minimum $Var[\hat{a}]$ (drawn from Tables 3.21 to 3.23) on the same table. Both versions agrees with each other by giving similar approximated values of $Var[\hat{a}]$ and the combination of κ and c. We also observe good conformity between simulated and theoretical results; the agreement is closer for larger r.

r	Q Version	GI Version
5	$\kappa = 0.86, c = 0.079$	$\kappa = 0.60, c = 0.147$
	Theoretical $Var\left[\hat{b}\right]_Q = 0.0052$ Practical $Var\left[\hat{b}\right] = 0.0058$	Theoretical $Var\left[\hat{b}\right]_Q = 0.0050$ Practical $Var\left[\hat{b}\right] = 0.0131$
10	$\kappa = 0.83, c = 0.156$	$\kappa = 0.69, c = 0.330$
	Theoretical $Var\left[\hat{b}\right]_Q = 0.0113$ Practical $Var\left[\hat{b}\right] = 0.0116$	Theoretical $Var\left[\hat{b}\right]_Q = 0.0106$ Practical $Var\left[\hat{b}\right] = 0.0247$

Table 3.26: Checking the accuracy of two versions of theoretical variance of attenuated moment estimator $\hat{b}$ for a mixture of two exponential distributions with $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical $Var\left[\hat{p}\right]_Q$	Theoretical $Var\left[\hat{p}\right]_{GI}$	Practical $Var\left[\hat{p}\right]$
2	0.1170 $(\kappa = 1.00, c = 0.030)$	0.1170 $(\kappa = 1.00, c = 0.030)$	0.0697 $(\kappa = 0.90, c = 0.010)$
5	0.0029 $(\kappa = 0.73, c = 0.058)$	0.0028 $(\kappa = 0.52, c = 0.120)$	0.0030 $(\kappa = 0.70, c = 0.060)$
10	0.0009 $(\kappa = 0.90, c = 0.265)$	0.0009 $(\kappa = 0.50, c = 0.209)$	0.0009 $(\kappa = 0.70, c = 0.100)$

Table 3.27: Theoretical and simulated minimum variance of attenuated moment estimator $\hat{p}$ given by the optimal combination of κ and c for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

In Table 3.25, we note the disagreement between $Var\left[\hat{b}\right]_Q$ and $Var\left[\hat{b}\right]_{GI}$ when $r = 5$ and $r = 10$. According to the GI-version, the best c is larger than 0.1 for $r = 5$ and $r = 10$. Our simulation work did not consider any c larger than 0.1, therefore another 10000 simulated samples were produced and each sample was estimated using the combination of κ and c suggested by both versions. The results are shown in Table 3.26. We find that the combination suggested by the Q-version does provide us with $Var\left[\hat{b}\right]$ smaller than the minimum we found as shown in Table 3.25, both for $r = 5$ and $r = 10$. On the other hand, the GI-version under-estimates $Var[b]$ for c larger than 0.1. Using the combination suggested by the GI-version, the practical $Var\left[\hat{b}\right]$ is actually twice more than the theoretical value suggested by the GI-version.

Similarly, in Table 3.27, both versions have a different optimal combination of κ and c for p when $r = 5$ and $r = 10$. For $r = 5$, $Var[p]_Q$ agrees with $Var[\hat{p}]$. We investigated the accuracy of $Var[\hat{p}]_{GI}$ by estimating 10000 simulated samples with $\kappa = 0.52$ and $c = 0.12$. The resulted $Var[\hat{p}]$ is 0.0106 instead of 0.0028. For $r = 10$, we estimated the simulated samples with $\kappa = 0.5$ and $c = 0.209$ and found that $Var[\hat{p}] = 0.0235$, instead of the theoretical value 0.0009. Hence, we further confirm that $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]_{GI}$ is only accurate when c is smaller than 0.1. We also used the Q-version $\kappa = 0.90$ and $c = 0.265$ to estimate 10000 simulated data sets and found the practical variance $Var[\hat{p}]$ as 0.0010, which is still marginally larger than the one we obtained ($Var[\hat{p}] = 0.0009$) with $\kappa = 0.7$ and $c = 0.1$ in our previous simulation work. However, this again proves that the Q-version of κ and c is reliable to return estimates with high precision.

Our investigation shows that the GI version provides a reasonable approximation of $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]_{GI}$ only for c smaller than 0.1. Encouragingly, simulated and theoretical values match up well for the Q-version. Although the Q-version of approximated $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ are reasonable, we do find the theoretical values differ marginally from the simulated value. There are two reasons behind this: first, the theoretical variances are approximations and only the first term of the Taylor expansion is used. Hence, there exists approximation errors in the computation of the theoretical variances. Second, there exists random errors in the simulation experiments. It is also worth mentioning that the optimal combination of κ and c minimises the variances of the estimators only when the sample size is large. Referring to Tables 3.21, 3.22 and 3.23, we find that the suggested combination does not return estimates with the best precision for small sample size.

To conclude, we suggest users employ the combination of κ and c suggested by the Q-version to estimate a two-component mixture exponential distribution with large sample size provided that they roughly know the separation between the two components. For r from 2 to 10, the optimal combinations of κ and c for r to estimate a, b and p are shown in Tables 3.28, 3.29 and 3.30 respectively. Note that the optimal c shown in these tables are only for mixtures with $a = 0.1$; for $a = 1$, the optimal c should be ten times the ones suggested in the tables. As seen in Table 3.28, the optimal fraction κ decreases gradually

r	κ	c	$Var\left[\hat{a}\right]_Q$
2	0.9718	0.0276	0.0003
3	0.8077	0.0335	9.49×10^{-5}
4	0.6920	0.0339	5.94×10^{-5}
5	0.6113	0.0328	4.61×10^{-5}
6	0.5527	0.0314	3.93×10^{-5}
7	0.5083	0.0299	3.51×10^{-5}
8	0.4735	0.0285	3.24×10^{-5}
9	0.4455	0.0272	3.04×10^{-5}
10	0.4225	0.0260	2.90×10^{-5}

Table 3.28: Optimal combination of κ and c and theoretical minimum variance of the attenuated moment estimator $\hat{a}$ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

r	κ	c	$Var\left[\hat{b}\right]_Q$
2	1.0174	0.0315	0.0056
3	0.9392	0.0495	0.0041
4	0.8902	0.0646	0.0045
5	0.8610	0.0790	0.0052
6	0.8440	0.0935	0.0062
7	0.8349	0.1084	0.0073
8	0.8311	0.1238	0.0085
9	0.8310	0.1398	0.0098
10	0.8333	0.1563	0.0113

Table 3.29: Optimal combination of κ and c and theoretical minimum variance of the attenuated moment estimator $\hat{b}$ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

from $\kappa = 0.9718$ when $r = 2$ to $\kappa = 0.4225$ when $r = 10$, while the optimal value of c differs marginally for different r; c is nearest to 0.03 for all r. $Var\left[\hat{a}\right]_Q$ decreases when the separation between the two components decreases. For the estimation of b, as presented in Table 3.29, the optimal κ decreases with a small margin when r decreases, especially for $r \geq 4$, the optimal fraction falls from $\kappa = 0.8902$ when $r = 4$ to $\kappa = 0.8333$ when $r = 10$; on the other hand, the optimal value of c increases when r increases. We also note that $Var\left[\hat{b}\right]_Q$ is lowest when $r = 3$ while $Var\left[\hat{b}\right]_Q$ actually increases for r larger than 3. From Table 3.30, we observe that the optimal κ for estimating p decreases when r increases from 2 up to 8; the optimal fraction increases from $\kappa = 0.7162$ when $r = 9$ to $\kappa = 0.9028$ when $r = 10$. Similar to a, the variance of p gets smaller when r increases.

Finally, we study the effect of κ and c on $Var\left[\hat{a}\right]_Q$, $Var\left[\hat{b}\right]_Q$ and $Var\left[\hat{p}\right]_Q$ for $r = 2$, 5 and 10, with $n_o = 1000$ in Figures 3.19, 3.20 and 3.21. For each r, we consider ten values of κ ranging from 0.1 to 1, with an increment of 0.1; for each κ, we plot $Var\left[\hat{\Theta}\right]_Q$ with respect to c. From Figure 3.19, we observe that combinations of a large fraction κ with a small

r	κ	c	$Var\,[\hat{p}]_Q$
2	0.9953	0.0297	0.1170
3	0.8725	0.0420	0.0138
4	0.7881	0.0501	0.0052
5	0.7325	0.0575	0.0029
6	0.6968	0.0656	0.0020
7	0.6774	0.0755	0.0015
8	0.6760	0.0900	0.0012
9	0.7162	0.1206	0.0010
10	0.9028	0.2649	0.0009

Table 3.30: Optimal combination of κ and c and theoretical minimum variance of the attenuated moment estimator $\hat{p}$ for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$.

attenuation c generally make $Var\,[\hat{a}]_Q$ small. When r increases, with a small attenuation c, the optimal κ decreases (also see Table 3.28). The optimal c is nearest to 0.03 for all r whereas $Var\,[\hat{a}]_Q$ is large for $c \geq 0.1$ in all cases considered here.

In Figure 3.20, we learn that a large fraction κ with an attenuation increases the precision of b. Unlike a, the optimal c increases for larger r. From the figure, we note that the minimum points shift to the right when r increases. This means that when the separation between the two components increases, the best combination is a large κ (which takes a value between 0.8 and 1) and a larger c (best c is 0.0315 for $r = 2$, 0.0790 for $r = 5$ and 0.1563 for $r = 10$).

Notably, the shape of the plots of $Var\,[p]_Q$ is to some extent different from the plots of $Var\,[a]_Q$ and $Var\,[b]_Q$. Our investigation shows that $Var\,[p]_Q$ has multiple local minimum and maximum points, as seen in Figure 3.21. However, the combinations of κ and c presented in Table 3.30 are the ones that give the global minimum of $Var\,[p]_Q$.

Undoubtedly, the optimal combination of κ and c depends on the separation between the two components in a mixture. In practice, r is unknown, how should we decide on the values of κ and c so that the resulted estimates are reliable? Of course, one can first estimate a raw data with the MLE and get an intuition about r from the ML estimates, so that a suitable combination of κ and c can be chosen from Tables 3.28 to 3.30. Alternatively, one can estimate a raw sample with different combinations of κ and c and choose the estimates that give the minimum asymptotic variance of estimator when they are substituted into (3.119).

We estimated the same simulated sample (with $n_o = 1000$) in Figure 3.13, which has true parameters $\Theta = (0.1, 0.5, 0.6)$, with ten combinations of κ and c, where κ is fixed as 0.8 and c is ranging from 0.01 to 0.1 with an increment of 0.01. We then substituted these ten sets of estimates into (3.119) and the resulted variances are plotted together with the asymptotic variances of estimator given by the true parameter values in Figure 3.22. Apart from b, the conformity between theory and practice is excellent, both for the optimal c and

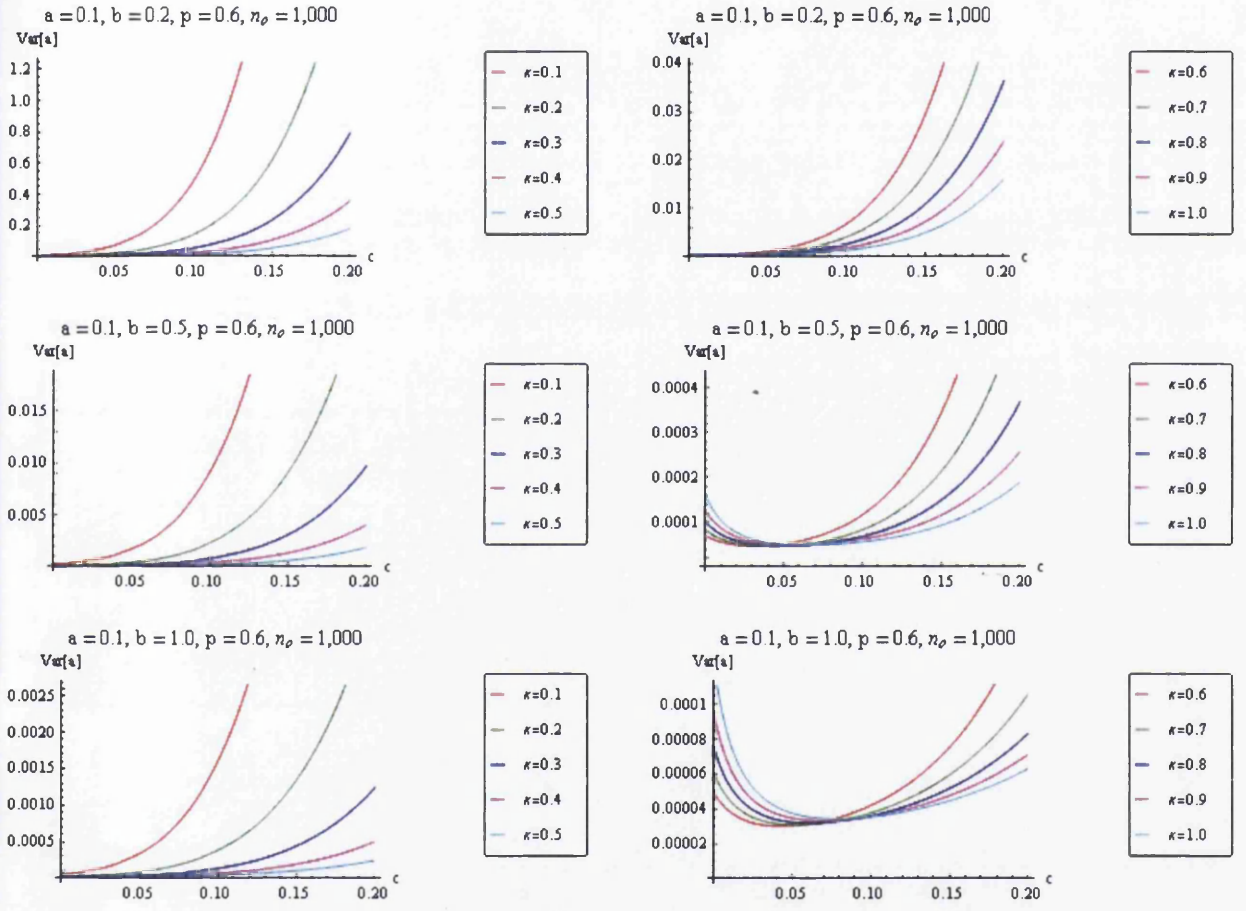


Figure 3.19: Plots of theoretical $Var\left[\hat{a}\right]_Q$ versus c for varying κ and r.

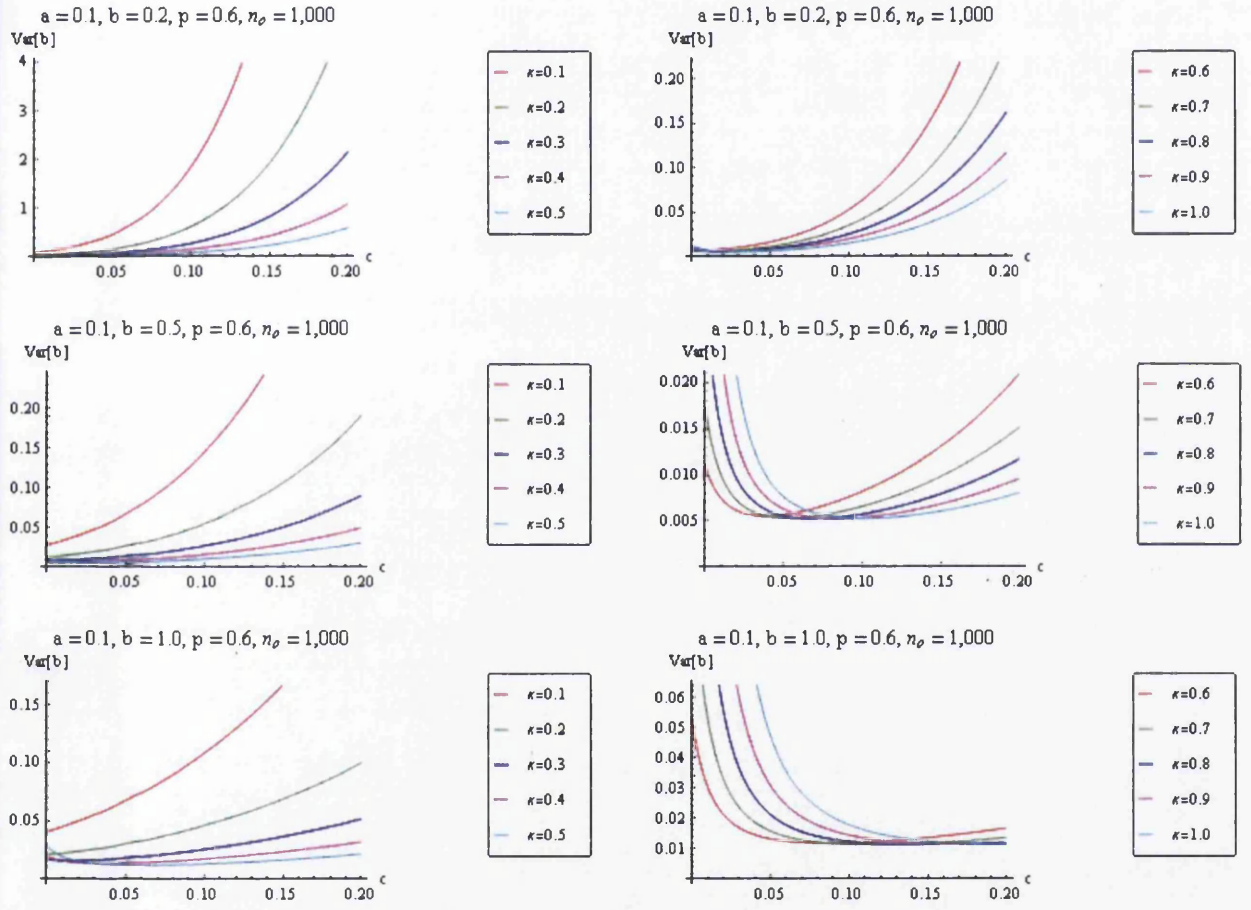


Figure 3.20: Plots of theoretical $Var\left[\hat{b}\right]_Q$ versus c for varying κ and r.

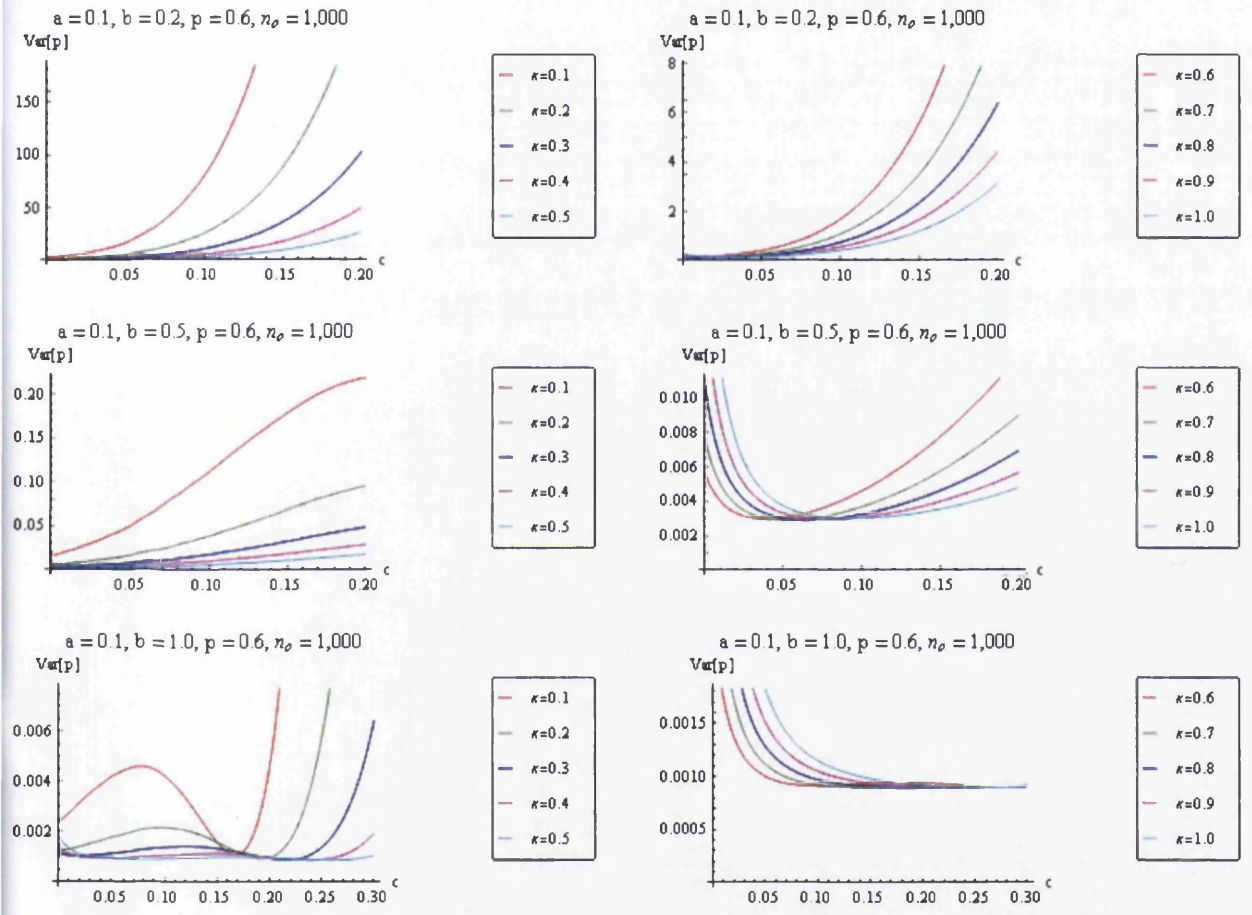


Figure 3.21: Plots of theoretical $Var\left[\hat{p}\right]_Q$ versus c for varying κ and r.

the variances. Therefore, in practice, the highest precise estimates should return the lowest $Var[\hat{a}]$, $Var\left[\hat{b}\right]$ and $Var[\hat{p}]$. Since the method of attenuated moments allows us to estimate the parameters quickly, we can consider as many combinations of κ and c as possible, and choose the set of estimates that gives the lowest variances of the estimators.

3.4.5 Discussion

The attenuated moment estimator is better than the method of moments and the method of fractional moments. With a small degree of attenuation, as we have seen from previous sections, the variance of attenuated moment estimator is relatively smaller than the variance of fractional moment estimator. We will show in the last section of this chapter that the performance of the attenuated moment estimator is indeed comparable to the MLE via EM algorithm. The main drawback of this new method is, like any moment-based method, it may yield unreasonable roots which are negative or in complex form.

3.5 The Method Based on an Appell Sequences

3.5.1 Introduction

Jalali (2005b) considered another method of parameter estimation which is similar to and inspired by the method of moments. The theoretical exposition in this subsection and the next follow his paper. Let $(\mu_1, ..., \mu_m)$ be a sequence of distinct scale parameters and $(p_1, ..., p_m)$ an exhaustive set of weights (i.e. $\sum_{i=1}^{m} p_j = 1$) (assume p_j's are positive). Suppose we have a sample arising from a mixture of m exponential distributions with the following PDF:

$$f(t) = \sum_{j=1}^{m} \frac{p_j}{\mu_j} \exp\left(-\frac{t}{\mu_j}\right). \tag{3.124}$$

Based on this sample, we want to estimate the scale and weight parameters. Let $h_0(t)$ be an integrable function. We define successive integrals of $h_0(t)$ as follows:

$$h_1(t) = \int_0^t h_0(\mu) d\mu,$$

and by recursion,

$$h_{m+1}(t) = \int_0^t h_m(\mu) d\mu. \tag{3.125}$$

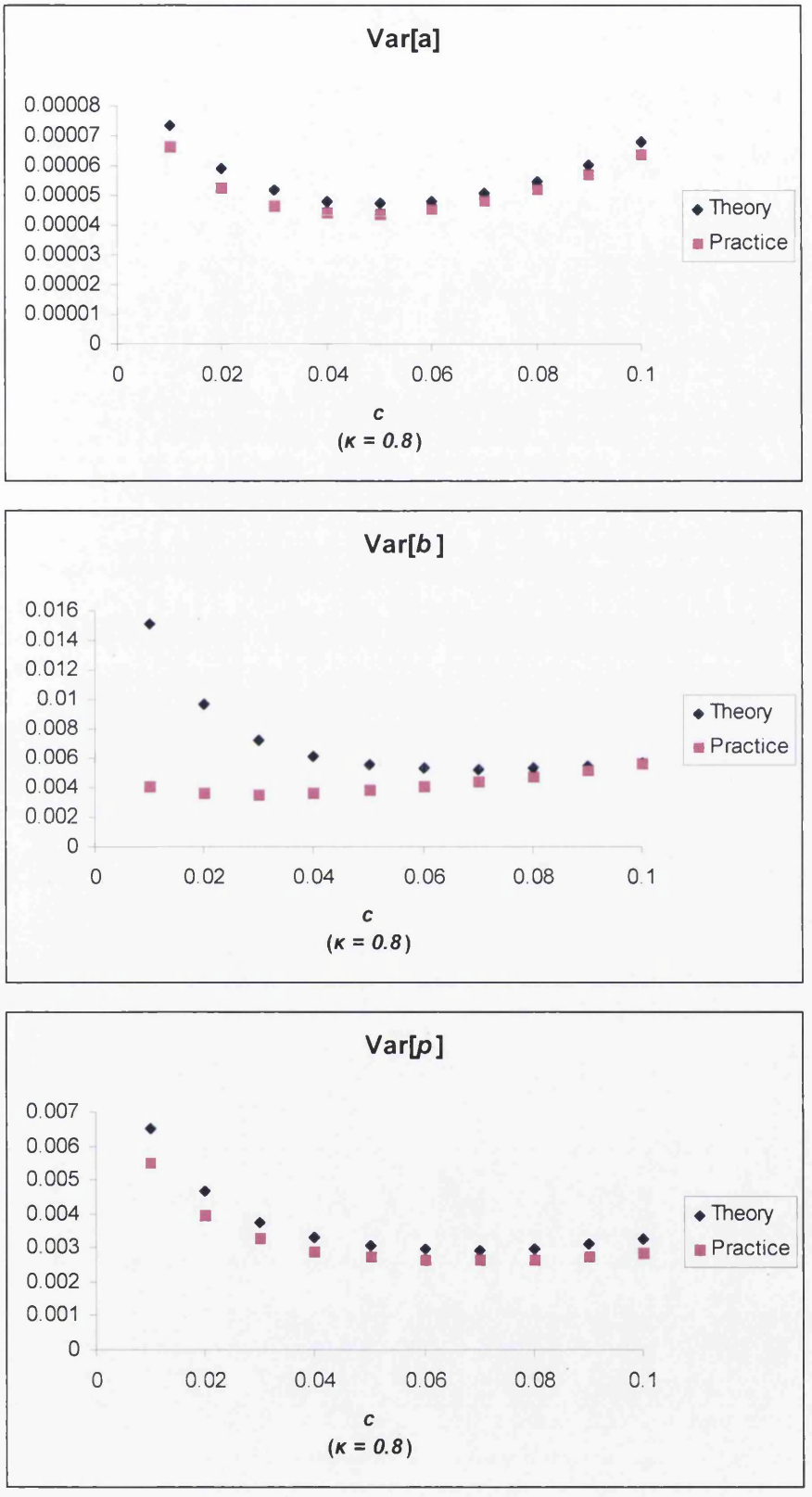


Figure 3.22: Asymptotic variance of the attenuated moment estimator given by true parameters and parameter estimates (estimated with $\kappa = 0.8$) versus c, based on a data set, consisting of 1000 observations, simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$.

We call the sequence of functions $\{h_k\}$ a basic Appell functions system based on $h_0(t)$. Now, consider the integral

$$\begin{aligned} I_{k+1}(\mu) &= {}^{def}\int_0^\infty \frac{h_{k+1}(t)}{\mu}\exp\left(-\frac{t}{\mu}\right)dt \\ &= \int_0^\infty \frac{\exp\left(-\frac{t}{\mu}\right)}{\mu}dt\int_0^t h_k(s)\,ds \\ &= \int_0^\infty h_k(s)\,ds\int_s^\infty \frac{\exp\left(-\frac{t}{\mu}\right)}{\mu}dt \\ &= \int_0^\infty h_k(s)\exp\left(-\frac{s}{\mu}\right)ds \\ &= \mu I_k(\mu). \end{aligned}$$

It follows by induction that

$$I_k(\mu) = \mu^k I_0(\mu) = \mu^k\int_0^\infty \frac{h_0(t)}{\mu}\exp\left(-\frac{t}{\mu}\right)dt. \tag{3.126}$$

(3.126) is the basis of our estimation. Define

$$\delta_k = \int_0^\infty h_k(t)\,f(t)\,dt = \int_0^\infty h_k(t)\left(\sum_{j=1}^m \frac{p_j}{\mu_j}\exp\left(-\frac{t}{\mu_j}\right)\right)dt = \sum_{j=1}^m p_j I_k(\mu_j) = \sum_{j=1}^m p_j I_0(\mu_j)\mu_j^k. \tag{3.127}$$

If we set $w_j = p_j I_0(\mu_j)$, then we have

$$\delta_k = \sum_{j=1}^m w_j\mu_j^k, \qquad k = 0, 1, ... \tag{3.128}$$

In the following we refer to these moment-like quantities as deltas. Now, let $\{t_i : i = 1, ..., n_o\}$ be a sample of size n_o from our population. Then the estimates of deltas are as follows:

$$\hat{\delta}_k = \frac{1}{n_o}\sum_{i=1}^{n_o} h_k(t_i). \tag{3.129}$$

If we find the first $2m$ estimates $\hat{\delta}_0, ..., \hat{\delta}_{2m-1}$, then we can find estimates for μ_j's as m roots of the following algebraic equation:

$$\det \begin{bmatrix} \hat{\delta}_0 & \hat{\delta}_1 & ... & \hat{\delta}_m \\ \hat{\delta}_1 & \hat{\delta}_2 & ... & \hat{\delta}_{m+1} \\ ... & ... & ... & ... \\ \hat{\delta}_{m-1} & \hat{\delta}_m & ... & \hat{\delta}_{2m-1} \\ 1 & u & ... & u^m \end{bmatrix} = 0. \tag{3.130}$$

We can next use the roots of this equation (estimates of μ_j's) and the first m $\hat{\delta}_j$'s to find estimates of w_j's as follows:

$$\begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \\ ... \\ \hat{w}_m \end{bmatrix} = \boldsymbol{V}^{-1} \begin{bmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ ... \\ \hat{\delta}_{m-1} \end{bmatrix}, \tag{3.131}$$

where V is the $m \times m$ Vandermonde matrix,

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 & ... & 1 \\ \hat{\mu}_1 & \hat{\mu}_2 & ... & \hat{\mu}_m \\ ... & ... & ... & ... \\ \hat{\mu}_1^{m-1} & \hat{\mu}_2^{m-1} & ... & \hat{\mu}_m^{m-1} \end{bmatrix}. \tag{3.132}$$

Next we shall find estimates of the weights as follows. First we find the auxiliary values

$$\gamma_j = \frac{\hat{w}_j}{I_0(\hat{\mu}_j)}, \tag{3.133}$$

and then we let

$$\gamma = \sum_{j=1}^{m} \gamma_j. \tag{3.134}$$

Finally,

$$\hat{p}_j = \frac{\gamma_j}{\gamma}. \tag{3.135}$$

As we have used $2m$ $\hat{\delta}$'s to find the estimates of the scales, we may as well use all of them in estimating the weights. To do this we define the following extended Vandermonde matrix for $\alpha \geq 2m - 1$.

$$\boldsymbol{V}_\alpha = \begin{bmatrix} 1 & 1 & ... & 1 \\ \hat{\mu}_1 & \hat{\mu}_2 & ... & \hat{\mu}_m \\ ... & ... & ... & ... \\ \hat{\mu}_1^{\alpha} & \hat{\mu}_2^{\alpha} & ... & \hat{\mu}_m^{\alpha} \end{bmatrix}. \tag{3.136}$$

This of course is a rectangular matrix which has m columns and $\alpha+1$ rows. Let

$$\Lambda = diag\left[\begin{array}{cccc} \lambda_0 & \lambda_1 & ..., & \lambda_\alpha \end{array}\right] \tag{3.137}$$

be a diagonal matrix with positive weights λ_j's on its main diagonal. When $\alpha > 2m-1$, $\boldsymbol{V}_\alpha$ has no ordinary inverse but equation (from (3.131))

$$\boldsymbol{V}_\alpha \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \\ \cdots \\ \hat{w}_m \end{bmatrix} = \begin{bmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \cdots \\ \hat{\delta}_\alpha \end{bmatrix} \tag{3.138}$$

is over-determined, and thus it has, in general, no solution. But we may ask what $\hat{w}_j$'s make the following weighted sum of squared errors minimum:

$$\left(\boldsymbol{V}_\alpha \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \\ \cdots \\ \hat{w}_m \end{bmatrix} - \begin{bmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \cdots \\ \hat{\delta}_\alpha \end{bmatrix}\right)^T \Lambda \left(\boldsymbol{V}_\alpha \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \\ \cdots \\ \hat{w}_m \end{bmatrix} - \begin{bmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \cdots \\ \hat{\delta}_\alpha \end{bmatrix}\right). \tag{3.139}$$

It is easy to see that the vector of weights which minimises the above quadratic form is

$$\begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \\ \cdots \\ \hat{w}_m \end{bmatrix} = (\boldsymbol{V}_\alpha^T \Lambda \boldsymbol{V}_\alpha)^{-1} \boldsymbol{V}_\alpha^T \Lambda \begin{bmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \cdots \\ \hat{\delta}_\alpha \end{bmatrix}. \tag{3.140}$$

We have to be careful that in (3.140), $\boldsymbol{V}_\alpha^T \Lambda \boldsymbol{V}_\alpha$ is likely to be a singular matrix. Now, the estimates of p_j's can be obtained as follows:

$$\hat{p}_j = \frac{\dfrac{\hat{w}_j}{I_0(\hat{\mu}_j)}}{\sum_{j=1}^{m} \dfrac{\hat{w}_j}{I_0(\hat{\mu}_j)}}. \tag{3.141}$$

3.5.2 Appell-Fourier Systems

Suppose we require $\alpha+1$ deltas. Then we need to look for the simplest functions of a family of functions with a zero of order α. If our family is that of the trigonometric functions, then such a function is $\sin^\alpha \varpi t$. So we let

$$h_\alpha(t) = \sin^\alpha \varpi t. \tag{3.142}$$

Now all the lower indexed h_k's can be found by differentiation. We need, however, the explicit form of these functions as we need to find estimates of deltas. In order to do this we write $\sin^{\alpha} \varpi t$ in terms of trigonometric functions of multiples of ϖt. There are two expressions depending on the parity of α:

$$\begin{aligned} \sin^{2A} \varpi t &= 2^{1-2A}\left[\sum_{i=0}^{A-1}(-1)^{A+i}\begin{pmatrix} 2A \\ i \end{pmatrix}\cos 2(A-i)\varpi t\right] + 2^{-2A}\begin{pmatrix} 2A \\ A \end{pmatrix} \quad (3.143) \\ \sin^{2A+1} \varpi t &= 2^{-2A}\sum_{i=0}^{A}(-1)^{A+i}\begin{pmatrix} 2A+1 \\ i \end{pmatrix}\sin(2A-2i+1)\varpi t \end{aligned}$$

(see Godoment (1969), page 182).

Clearly,

$$\int_0^{\infty} \frac{\cos \varpi t}{\mu} \exp\left(-\frac{t}{\mu}\right) dt = \frac{1}{1+\varpi^2\mu^2}$$

and

$$\int_0^{\infty} \frac{\sin \varpi t}{\mu} \exp\left(-\frac{t}{\mu}\right) dt = \frac{\varpi\mu}{1+\varpi^2\mu^2}.$$

Recalling from (3.126) that

$$I_0(\mu) = \mu^{-m} I_m(\mu),$$

and we have

$$I_0(\mu) = \begin{cases} \dfrac{2}{(2\mu)^{\alpha}}\left\{\displaystyle\sum_{i=0}^{A-1}\frac{(-1^{A+i}\begin{pmatrix} 2A \\ i \end{pmatrix}}{1+[2(A-i)\varpi\mu]^2} + \frac{1}{2}\begin{pmatrix} 2A \\ A \end{pmatrix}\right\} & \text{if} \quad \alpha = 2A \\ \dfrac{2}{(2\mu)^{\alpha}}\left\{\displaystyle\sum_{i=0}^{A}\frac{(-1)^{A+i}\begin{pmatrix} 2A+1 \\ i \end{pmatrix}[(2A-2i+1)\varpi\mu]}{1+[(2A-2i+1)\varpi\mu]^2}\right\} & \text{if} \quad \alpha = 2A+1. \end{cases} \quad (3.144)$$

Now we can find

$$w_j = p_j I_0(\mu_j). \quad (3.145)$$

$\hat{\delta}_j$'s can be obtained as follows:

$$\hat{\delta}_k = \frac{1}{n_o}\sum_{i=1}^{n_o}\sin^k \varpi t_i. \quad (3.146)$$

We then can find the estimates of μ_j's, w_j's and p_j's following the procedure from (3.130) to (3.141). Note that in this method we are free to choose ϖ and $\lambda_0, ..., \lambda_{2m-2}$. As powers

of ϖ appear successively in the expression for $h_k(t)$ the obvious candidates for ϖ are 1, and any other number whose $(2m-1)^{th}$ power is neither too small nor too large.

Our simulation work considers the performance of the this method when it is used to estimate a mixture of two exponential distributions. In the followings, we presents the procedures of the estimation for $\alpha = 3$, 4 and 5.

$\alpha = 3$

In this particular case of $m = 2$, we have

$$\begin{aligned} h_3(t) &= \sin^3 \varpi t, \\ h_2(t) &= 3\varpi \sin^2 \varpi t \cos \varpi t, \\ h_1(t) &= 3\varpi^2 \sin \varpi t \left(2\cos^2 \varpi t - \sin^2 \varpi t\right), \\ h_0(t) &= 3\varpi^3 \cos \varpi t \left(2\cos^2 \varpi t - 7\sin^2 \varpi t\right). \end{aligned} \tag{3.147}$$

Using formula (3.143) where $A = 1$, (3.147) becomes

$$\begin{aligned} h_3(t) &= \sin^3 \varpi t \\ &= 2^{-2}\left[(-1)\begin{pmatrix} 3 \\ 0 \end{pmatrix} \sin 3\varpi t + (-1)^2 \begin{pmatrix} 3 \\ 1 \end{pmatrix} \sin\left(2(1) - 2(1) + 1\right)\varpi t\right] \\ &= \frac{1}{4}\left[3\sin \varpi t - \sin 3\varpi t\right], \end{aligned} \tag{3.148}$$

$$h_2(t) = \frac{\varpi}{4}\left[3\cos \varpi t - 3\cos 3\varpi t\right], \tag{3.149}$$

$$h_1(t) = \frac{-\varpi^2}{4}\left[3\sin \varpi t - 9\sin 3\varpi t\right], \tag{3.150}$$

$$h_0(t) = -\frac{3\varpi^3}{4}\left[\cos \varpi t - 9\cos 3\varpi t\right]. \tag{3.151}$$

Using these functions we can find the estimates of the first four deltas given a data set consisting of n_o observations as follows:

$$\begin{aligned} \hat{\delta}_0 &= \frac{1}{n_o}\sum_{i=1}^{n_o}\left(-\frac{3\varpi^3}{4}\left[\cos \varpi t_i - 9\cos 3\varpi t_i\right]\right), \\ \hat{\delta}_1 &= \frac{1}{n_o}\sum_{i=1}^{n_o}\left(\frac{-\varpi^2}{4}\left[3\sin \varpi t_i - 9\sin 3\varpi t_i\right]\right), \\ \hat{\delta}_2 &= \frac{1}{n_o}\sum_{i=1}^{n_o}\left(\frac{\varpi}{4}\left[3\cos \varpi t_i - 3\cos 3\varpi t_i\right]\right), \\ \hat{\delta}_3 &= \frac{1}{n_o}\sum_{i=1}^{n_o}\left(\frac{1}{4}\left[3\sin \varpi t_i - \sin 3\varpi t_i\right]\right). \end{aligned} \tag{3.152}$$

Then we should find the roots of the determinantal equation

$$\det \begin{bmatrix} \hat{\delta}_0 & \hat{\delta}_1 & \hat{\delta}_2 \\ \hat{\delta}_1 & \hat{\delta}_2 & \hat{\delta}_3 \\ 1 & u & u^2 \end{bmatrix} = 0, \tag{3.153}$$

which are the estimates of μ_1 and μ_2, and hence

$$\begin{aligned} \hat{a} &= \frac{1}{\hat{\mu}_1}, \\ \hat{b} &= \frac{1}{\hat{\mu}_2}. \end{aligned} \tag{3.154}$$

From (3.140) we can also find estimates of w_1 and w_2. We note that, according to (3.144),

$$I_0(\mu_j) = \frac{3\varpi\mu}{4\mu_j^3}\left[\frac{1}{1+\varpi^2\mu_j^2} - \frac{1}{1+9\varpi^2\mu_j^2}\right] = \frac{6\varpi^3}{(1+\varpi^2\mu_j^2)(1+9\varpi^2\mu_j^2)}. \tag{3.155}$$

This enables us to find the estimate of p_j using (3.141).

$\alpha = 4$

When $\alpha = 4$ and $m = 2$, we need five Appell sample moments $\hat{\delta}_k = \frac{1}{n_o}\sum_{i=1}^{n_o} h_k(t_i)$ for $k = 0, 1, ..., 4$ where

$$\begin{aligned} h_4(t) &= \frac{1}{8}\cos 4\omega t - \frac{1}{2}\cos 2\omega t + \frac{3}{8}, \\ h_3(t) &= \omega\left[-\frac{1}{2}\sin 4\omega t \sin 2\omega t\right], \\ h_2(t) &= 2\varpi^2\left[-\cos 4\omega t \cos 2\omega t\right], \\ h_1(t) &= 4\varpi^3\left[2\sin 4\omega t - \sin 2\omega t\right], \\ h_0(t) &= 8\varpi^4\left[4\cos 4\omega t - \cos 2\omega t\right], \end{aligned} \tag{3.156}$$

and

$$I_0(\mu_j) = \frac{24\varpi^4}{(1+4\varpi^2\mu_j^2)(1+16\varpi^2\mu_j^2)}. \tag{3.157}$$

Upon substituting the observed values of $\hat{\delta}_0$, $\hat{\delta}_1$, $\hat{\delta}_2$ and $\hat{\delta}_3$ given by (3.156) into the determinantal equation (3.130), the estimates of μ_1 and μ_2 are then given by the roots of the quadratic equation; $\hat{a}$ and $\hat{b}$ are just the inverses of $\hat{\mu}_1$ and $\hat{\mu}_2$ respectively. To fully utilise the five $\hat{\delta}_k$'s we have here, we put all of them in (3.140) to estimate w_1 and w_2. Given $\hat{w}_1$, $\hat{w}_2$, $I_0(\hat{\mu}_1)$ and $I_0(\hat{\mu}_2)$ (given by (3.157)), we obtain an estimate of p by following (3.141).

$\alpha = 5$

When $\alpha = 5$ and $m = 2$, we make use of six Appell sample moments $\hat{\delta}_k = \frac{1}{n_o}\sum_{i=1}^{n_o} h_k(t_i)$ for $k = 0, 1, ..., 5$ where

$$\begin{aligned}
h_5(t) &= \frac{1}{16}\sin 5\omega t - \frac{5}{16}\sin 3\omega t + \frac{5}{8}\sin \omega t, \qquad (3.158)\\
h_4(t) &= \frac{5}{8}\omega\left[\frac{1}{2}\cos 5\omega t - \frac{3}{2}\cos 3\omega t + \cos \omega t\right],\\
h_3(t) &= \frac{5}{8}\omega^2\left[-\frac{5}{2}\sin 5\omega t + \frac{9}{2}\sin 3\omega t - \sin \omega t\right],\\
h_2(t) &= \frac{5}{8}\omega^3\left[-\frac{25}{2}\cos 5\omega t + \frac{27}{2}\cos 3\omega t - \cos \omega t\right],\\
h_1(t) &= \frac{5}{8}\omega^4\left[\frac{125}{2}\sin 5\omega t - \frac{81}{2}\sin 3\omega t + \sin \omega t\right],\\
h_0(t) &= \frac{5}{8}\omega^5\left[\frac{625}{2}\cos 5\omega t - \frac{243}{2}\cos 3\omega t + \cos \omega t\right],
\end{aligned}$$

and

$$I_0(\mu_j) = \frac{120\varpi^5}{(1+25\varpi^2\mu_j^2)(1+9\varpi^2\mu_j^2)(1+\varpi^2\mu_j^2)}. \qquad (3.159)$$

Like before, we substitute the first four $\hat{\delta}_k$'s given by (3.158) into the determinantal equation (3.130). $\hat{\mu}_1$ is the larger root of the quadratic equation, whereas $\hat{\mu}_2$ is given by the smaller root. We make full use of the six $\hat{\delta}_k$'s by substituting them into (3.140) to estimate w_1 and w_2. We substitute $\hat{\mu}_1$ and $\hat{\mu}_2$ into (3.159) to obtain $I_0(\hat{\mu}_1)$ and $I_0(\hat{\mu}_2)$, together with $\hat{w}_1$ and $\hat{w}_2$, we find $\hat{p}$ from (3.141).

3.5.3 Simulation Results

An investigation of the performance characteristics of the method using Appell sequences over the mixture of two exponential distributions is reported in this section. We estimate mixture distributions with different levels of separation between the components, varying from small ($r = 2$) over medium ($r = 5$) to large ($r = 10$). The choice of ϖ is crucial in order to obtain reasonable estimates; therefore, we have attempted to find the optimal value of ϖ which gives the best estimates of the parameters. We compared simulation results with different value of λ_0, λ_1 to λ_α in (3.137) and found that any change in the values of $\mathbf{\Lambda}$ makes no difference to our results. Therefore, we set all λ's as 1 for our parameter estimation with this method.

Now, let us study the simulation results of the method using Appell Sequences from Tables 3.31 to 3.39. Like before, we consider three degrees of separation: $r = (2, 5, 10)$ for samples with sizes $n_o = (10, 15, 20, 50, 1000)$. We consider nineteen values of ω, at which the first ten ω's ranging from 0.01 to 0.1, with an increment of 0.01 and the second nine ω's varying from 0.2 to 1 with an increment of 0.1. For each of these combinations, we

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	2.31×10^{-6} (0.03)	1.08×10^{-6} (0.01)	2.01×10^{-7} (0.04)	6.56×10^{-6} (0.04)	6.47×10^{-7} (0.01)
$\left(\bar{\hat{b}} - b\right)^2$	0.0048 (0.8)	0.0039 (0.03)	0.0013 (0.04)	5.27×10^{-5} (0.06)	0.0018 (0.08)
$(\bar{\hat{p}} - p)^2$	0.0006 (0.04)	6.76×10^{-5} (0.7)	0.0021 (0.8)	0.0004 (0.02)	0.0002 (0.02)
$Var[\hat{a}]$	0.0037 (0.01)	0.0027 (0.01)	0.0023 (0.01)	0.0016 (0.01)	0.0003 (0.01)
$Var\left[\hat{b}\right]$	275 (0.04)	84 (0.01)	68 (0.03)	172 (0.06)	21 (0.05)
$Var[\hat{p}]$	0.4155 (0.01)	0.5516 (0.01)	0.4615 (0.01)	0.2599 (0.01)	0.0765 (0.02)
$MSE[\hat{a}]$	0.0037 (0.01)	0.0027 (0.01)	0.0023 (0.01)	0.0016 (0.01)	0.0004 (0.01)
$MSE\left[\hat{b}\right]$	275 (0.04)	84 (0.01)	69 (0.03)	172 (0.06)	21 (0.05)
$MSE[\hat{p}]$	0.4262 (0.01)	0.5590 (0.01)	0.4679 (0.01)	0.2606 (0.01)	0.0767 (0.02)

Table 3.31: Performance of the method of Appell moments (with $\alpha = 3$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$ for different sample size n_o.

simulate 10000 mixture exponential data sets and estimate them with the method using Appell Sequences. For each combination of r and n_o, we recorded the minimum measures of error, namely the average bias^2, the variance and the mean squared error, and their counterpart ω in a bracket in the tables below. Tables 3.31 to 3.33 evaluate the minimum estimation error of the Appell Sequences estimator with $\alpha = 3$ (abbreviated as AP_3); Tables 3.34 to 3.36 present the best simulation result of the Appell Sequences estimator using $\alpha = 4$ (abbreviated as AP_4); whereas from Tables 3.37 to 3.39, we see the minimum measures of error of the Appell Sequences estimator with $\alpha = 5$ (abbreviated as AP_5).

Let us first focus on the performance of the Appell Sequences estimator with $\alpha = 3$. From Table 3.31, we can see that for $r = 2$, the ideal candidate of ω for estimating a is 0.01 for all sample sizes considered here. When $n_o = 1000$, the minimum variance of estimator of a is $Var[\hat{a}] = 0.0003$ with $\omega = 0.01$. The $Var\left[\hat{b}\right]$ is rather large and we cannot conclude the best value of ω given our simulation results. For p, the best ω is 0.01 for $n_o \leq 50$ and $\omega = 0.02$ returns the minimum $Var[\hat{p}] = 0.0765$ when $n_o = 1000$.

For an exponential mixture distribution with $r = 5$, we observe, from Table 3.32, that the optimal ω for a is 0.01 for small samples ($n_o \leq 50$); whereas the best ω is 0.04 when $n_o = 1000$, giving $Var[\hat{a}] = 7.01 \times 10^{-5}$. $Var[b]$ is large for $n_o \leq 50$. Excitingly, using $\omega = 0.08$ on samples with $n_o = 1000$ we obtained a small $Var\left[\hat{b}\right] = 0.0173$, much smaller than the $Var\left[\hat{b}\right]$ for $r = 2$. For $n_o \leq 50$, the best ω for p is either 0.01 or 0.02; whereas the

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	2.48×10^{-5} (0.07)	0.0002 (0.09)	8.93×10^{-5} (0.04)	2.20×10^{-7} (0.05)	2.18×10^{-10} (0.04)
$(\bar{\hat{b}} - b)^2$	0.0417 (0.02)	0.0007 (0.07)	0.0019 (0.09)	0.0005 (0.04)	7.47×10^{-6} (0.2)
$(\bar{\hat{p}} - p)^2$	4.69×10^{-5} (0.04)	0.0003 (0.8)	0.0010 (0.6)	0.0004 (0.06)	5.76×10^{-8} (0.04)
$Var[\hat{a}]$	0.0063 (0.01)	0.0038 (0.01)	0.0030 (0.01)	0.0013 (0.01)	7.01×10^{-5} (0.04)
$Var\left[\hat{b}\right]$	236 (0.09)	657 (0.1)	243 (0.04)	283 (0.05)	0.0173 (0.08)
$Var[\hat{p}]$	0.2855 (0.01)	0.1563 (0.01)	0.1792 (0.01)	0.0795 (0.02)	0.0050 (0.06)
$MSE[\hat{a}]$	0.0070 (0.01)	0.0042 (0.01)	0.0033 (0.02)	0.0015 (0.02)	7.01×10^{-5} (0.04)
$MSE\left[\hat{b}\right]$	237 (0.09)	657 (0.1)	244 (0.04)	283 (0.05)	0.0182 (0.08)
$MSE[\hat{p}]$	0.2985 (0.01)	0.1655 (0.01)	0.1893 (0.01)	0.0850 (0.02)	0.0049 (0.06)

Table 3.32: Performance of the method of Appell moments (with $\alpha = 3$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$ for different sample size n_o.

optimal ω for p is 0.06 when $n_o = 1000$, giving the minimum $Var[\hat{p}] = 0.0049$.

For $r = 10$, the optimal ω is either 0.01 or 0.02 for $n_o \leq 50$; whereas the best $\omega = 0.04$ returns the minimum $Var[\hat{a}] = 5.05 \times 10^{-5}$. For b, again, the variances are high for samples with small sizes ($n_o \leq 50$); When $n_o = 1000$, $\omega = 0.2$ has the minimum $Var\left[\hat{b}\right] = 0.0444$, which is indeed larger than the $Var\left[\hat{b}\right]$ for $r = 5$ with the same sample size. Similar to $r = 5$, the best ω for estimating p is either 0.01 or 0.02 when sample sizes are small ($n_o \leq 50$). Using $\omega = 0.1$ for samples with size $n_o = 1000$, one can obtain the minimum $Var[\hat{p}] = 0.0016$.

Next, we interpret the simulation results of the Appell Sequences estimator with $\alpha = 4$. For $r = 2$ (see Table 3.34), the best ω for estimating a is $\omega = 0.01$ for any sample size. The minimum $Var\left[\hat{b}\right]$'s we obtained are large, even for a sample as large as $n_o = 1000$; however, in order to get the best attainable estimate of b, one should use any $\omega \leq 0.03$. Unlike the other estimators, the Appell Sequences estimator $\hat{p}$ is unsatisfactory for small samples; even when $n_o = 1000$, $Var[\hat{p}] = 0.6143$ is about 8 times larger than the one given by the AP_3 (where $Var[\hat{p}] = 0.0765$).

Let us study the estimation results for $r = 5$ and $r = 10$ in Tables 3.35 and 3.36; they are similar to the ones when $r = 2$. For both r, the difference between the $Var[\hat{a}]$ of the AP_3 and the AP_4 is marginally small. Nevertheless, $Var\left[\hat{b}\right]$ of the AP_4 for $n_o = 1000$ is

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{\bar{a}} - a)^2$	4.46×10^{-5} (0.06)	0.0006 (0.1)	0.0002 (0.05)	2.46×10^{-8} (0.06)	3.01×10^{-10} (0.05)
$(\hat{\bar{b}} - b)^2$	0.0534 (0.6)	0.0079 (0.07)	0.0205 (0.2)	0.0420 (0.05)	0.0012 (0.2)
$(\hat{\bar{p}} - p)^2$	5.20×10^{-5} (0.09)	0.0005 (0.3)	4.55×10^{-5} (0.05)	8.76×10^{-5} (0.05)	1.19×10^{-8} (0.1)
$Var[\hat{a}]$	0.0097 (0.01)	0.0049 (0.02)	0.0034 (0.01)	0.0012 (0.01)	5.05×10^{-5} (0.04)
$Var[\hat{b}]$	615 (0.05)	279 (0.04)	963 (0.02)	646 (0.04)	0.0444 (0.2)
$Var[\hat{p}]$	0.1608 (0.01)	0.1011 (0.01)	0.0866 (0.02)	0.0445 (0.01)	0.0016 (0.1)
$MSE[\hat{a}]$	0.0116 (0.01)	0.0057 (0.02)	0.0041 (0.01)	0.0013 (0.04)	5.05×10^{-5} (0.04)
$MSE[\hat{b}]$	617 (0.05)	280 (0.04)	965 (0.02)	647 (0.04)	0.0455 (0.2)
$MSE[\hat{p}]$	0.1756 (0.01)	0.1192 (0.01)	0.0998 (0.02)	0.0514 (0.05)	0.0016 (0.1)

Table 3.33: Performance of the method of Appell moments (with $\alpha = 3$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$ for different sample size n_o.

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{\bar{a}} - a)^2$	1.24×10^{-5} (0.06)	2.20×10^{-8} (0.01)	4.57×10^{-6} (0.01)	2.04×10^{-6} (0.03)	3.44×10^{-6} (0.05)
$(\hat{\bar{b}} - b)^2$	0.0088 (0.07)	2.08×10^{-6} (0.05)	2.90×10^{-6} (0.3)	0.0006 (0.6)	1.60×10^{-6} (0.03)
$(\hat{\bar{p}} - p)^2$	0.0004 (0.06)	0.0119 (0.06)	0.0008 (0.04)	0.0038 (0.04)	0.0014 (0.05)
$Var[\hat{a}]$	0.0040 (0.01)	0.0028 (0.01)	0.0023 (0.01)	0.0016 (0.01)	0.0004 (0.01)
$Var[\hat{b}]$	396 (0.03)	229 (0.01)	84 (0.02)	315 (0.03)	13 (0.03)
$Var[\hat{p}]$	63 (0.01)	53 (0.02)	68 (0.05)	66 (0.3)	0.6143 (0.5)
$MSE[\hat{a}]$	0.0040 (0.01)	0.0028 (0.01)	0.0023 (0.01)	0.0016 (0.01)	0.0004 (0.01)
$MSE[\hat{b}]$	396 (0.03)	229 (0.01)	84 (0.02)	315 (0.03)	13 (0.03)
$MSE[\hat{p}]$	63 (0.01)	53 (0.02)	68 (0.05)	66 (0.3)	0.6163 (0.5)

Table 3.34: Performance of the method of Appell moments (with $\alpha = 4$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\overline{\hat{a}} - a)^2$	0.0006 (0.07)	7.89×10^{-5} (0.02)	0.0002 (0.02)	2.34×10^{-6} (0.03)	4.62×10^{-8} (0.02)
$\left(\overline{\hat{b}} - b\right)^2$	0.0003 (0.02)	0.0916 (0.04)	0.0050 (0.08)	0.0076 (0.05)	0.0018 (0.02)
$(\overline{\hat{p}} - p)^2$	0.0071 (0.04)	3.09×10^{-7} (0.09)	0.0002 (0.08)	5.61×10^{-5} (0.07)	7.38×10^{-5} (0.08)
$Var[\hat{a}]$	0.0062 (0.01)	0.0039 (0.01)	0.0029 (0.01)	0.0014 (0.01)	7.61×10^{-5} (0.02)
$Var\left[\hat{b}\right]$	207 (0.09)	224 (0.04)	402 (0.03)	211 (0.02)	0.0774 (0.05)
$Var[\hat{p}]$	128 (0.01)	72 (0.4)	140 (0.07)	6 (0.3)	0.1913 (0.5)
$MSE[\hat{a}]$	0.0069 (0.01)	0.0043 (0.01)	0.0031 (0.01)	0.0015 (0.01)	7.62×10^{-5} (0.02)
$MSE\left[\hat{b}\right]$	208 (0.09)	224 (0.04)	402 (0.03)	211 (0.02)	0.0800 (0.05)
$MSE[\hat{p}]$	128 (0.01)	72 (0.4)	140 (0.07)	6 (0.3)	0.2616 (0.5)

Table 3.35: Performance of the method of Appell moments (with $\alpha = 4$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$ for different sample size n_o.

greatly reduced, compared to $Var\left[\hat{b}\right]$'s from small samples. The estimation of p remains poor; it is so far the only estimator that returns extremely large $Var[\hat{p}]$ when the sample size is small. For samples with $r = 5$ and $n_o = 1000$, $Var[\hat{p}] = 0.1913$ is 39 times larger than the one given by the AP_3 ($= 0.0049$); whereas for samples with $r = 10$ and $n_o = 1000$, $Var[\hat{p}] = 0.0571$ is 35 times bigger than the one provided by the AP_3 ($= 0.0016$). Given that the AP_4 returns bad estimates of b and p, it is implausible for the parameter estimation of a two-component exponential mixture model.

We now examine the performance of the Appell Sequences estimator with $\alpha = 5$. From Tables 3.37, 3.38 and 3.39, we observe that, for all r considered here, $Var[\hat{a}]$ is relatively large compared to the ones provided by the AP_3 and AP_4 when the sample size is $n_o = 10$ and 15. Strangely, this estimator has a weak performance in estimating b and p for samples of large sizes ($n_o = 1000$). Its $Var\left[\hat{b}\right]$ and $Var[\hat{p}]$ are relatively larger than the ones given by the AP_3 and AP_4.

To summarise, the AP_3 outperforms AP_4 and AP_5 because it provides reasonable estimates of a and p. Indeed, $Var\left[\hat{b}\right]$ of the AP_3 is small when the sample size is large ($n_o = 1000$) and the separation between two populations is large ($r = 5$ and $r = 10$).

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	0.0004 (0.09)	0.0002 (0.03)	6.42×10^{-7} (0.06)	7.10×10^{-6} (0.03)	3.04×10^{-9} (0.03)
$(\bar{\hat{b}} - b)^2$	0.0100 (0.8)	0.5752 (0.1)	0.0041 (0.06)	0.0307 (0.08)	0.0079 (0.1)
$(\bar{\hat{p}} - p)^2$	0.0002 (0.2)	0.0050 (0.6)	1.52×10^{-5} (0.07)	0.0032 (0.09)	0.0079 (0.06)
$Var[\hat{a}]$	0.0318 (0.01)	0.0049 (0.01)	0.0035 (0.01)	0.0013 (0.01)	5.75×10^{-5} (0.02)
$Var[\hat{b}]$	282 (0.07)	421 (0.01)	444 (0.05)	2253 (0.08)	0.1887 (0.1)
$Var[\hat{p}]$	54 (0.3)	258 (0.06)	11 (0.3)	0.9649 (0.3)	0.0571 (0.2)
$MSE[\hat{a}]$	0.0334 (0.01)	0.0059 (0.01)	0.0041 (0.01)	0.0015 (0.01)	5.76×10^{-5} (0.02)
$MSE[\hat{b}]$	284 (0.07)	424 (0.01)	445 (0.05)	2253 (0.08)	0.1966 (0.1)
$MSE[\hat{p}]$	54 (0.3)	259 (0.06)	11 (0.3)	1.0132 (0.3)	0.0697 (0.2)

Table 3.36: Performance of the method of Appell moments (with $\alpha = 4$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$ for different sample size n_o.

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	1.15×10^{-7} (0.01)	8.48×10^{-7} (0.01)	8.40×10^{-6} (0.01)	1.71×10^{-6} (0.02)	1.21×10^{-5} (0.01)
$(\bar{\hat{b}} - b)^2$	0.0009 (0.3)	0.0042 (0.9)	0.0113 (0.3)	0.0003 (0.5)	0.0038 (0.01)
$(\bar{\hat{p}} - p)^2$	0.0034 (0.06)	0.0096 (0.1)	0.0248 (0.05)	6.01×10^{-5} (0.05)	0.0020 (0.08)
$Var[\hat{a}]$	0.0399 (0.01)	0.0139 (0.01)	0.0024 (0.01)	0.0017 (0.01)	0.0004 (0.01)
$Var[\hat{b}]$	307 (0.02)	215 (0.03)	143 (0.05)	309 (0.04)	18 (0.02)
$Var[\hat{p}]$	1.1156 (0.01)	0.1869 (0.01)	0.1808 (0.01)	0.1943 (0.01)	8 (0.01)
$MSE[\hat{a}]$	0.0399 (0.01)	0.0139 (0.01)	0.0024 (0.01)	0.0017 (0.01)	0.0004 (0.01)
$MSE[\hat{b}]$	307 (0.02)	215 (0.03)	143 (0.05)	309 (0.04)	18 (0.02)
$MSE[\hat{p}]$	1.2142 (0.01)	0.2945 (0.01)	0.2990 (0.01)	0.3191 (0.01)	8 (0.01)

Table 3.37: Performance of the method of Appell moments (with $\alpha = 5$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	0.0005 (0.01)	3.02×10^{-5} (0.06)	0.0002 (0.01)	9.50×10^{-6} (0.02)	2.16×10^{-9} (0.02)
$\left(\bar{\hat{b}} - b\right)^2$	0.0111 (0.03)	0.0307 (0.01)	0.0040 (0.03)	0.0375 (0.03)	0.0054 (0.06)
$(\bar{\hat{p}} - p)^2$	0.0103 (0.04)	0.0002 (0.06)	0.0037 (0.5)	0.0008 (0.07)	0.0024 (1)
$Var[\hat{a}]$	0.2122 (0.01)	0.2213 (0.02)	0.0033 (0.01)	0.0015 (0.01)	8.94×10^{-5} (0.01)
$Var\left[\hat{b}\right]$	444 (0.06)	466 (0.08)	497 (0.08)	172 (0.04)	1.0180 (0.04)
$Var[\hat{p}]$	0.1788 (0.01)	0.1790 (0.01)	0.1510 (0.01)	0.1031 (0.01)	3 (0.01)
$MSE[\hat{a}]$	0.2127 (0.01)	0.2219 (0.02)	0.0035 (0.01)	0.0016 (0.01)	8.94×10^{-5} (0.01)
$MSE\left[\hat{b}\right]$	444 (0.06)	467 (0.08)	498 (0.08)	172 (0.04)	1.0254 (0.04)
$MSE[\hat{p}]$	0.2565 (0.01)	0.2786 (0.01)	0.2639 (0.01)	0.2695 (0.01)	4 (0.01)

Table 3.38: Performance of the method of Appell moments (with $\alpha = 5$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	2.71×10^{-6} (0.03)	0.0003 (0.03)	0.0003 (0.03)	1.53×10^{-5} (0.02)	2.84×10^{-8} (0.02)
$\left(\bar{\hat{b}} - b\right)^2$	0.0187 (0.09)	0.0051 (0.04)	0.0095 (0.6)	0.2502 (0.04)	0.0044 (0.1)
$(\bar{\hat{p}} - p)^2$	0.0005 (0.07)	0.0089 (0.2)	7.12×10^{-5} (1)	0.0002 (0.05)	0.0046 (0.6)
$Var[\hat{a}]$	0.0184 (0.01)	0.2798 (0.01)	0.0040 (0.01)	0.0014 (0.01)	6.67×10^{-5} (0.02)
$Var\left[\hat{b}\right]$	94 (0.06)	424 (0.08)	316 (0.05)	825 (0.03)	11 (0.07)
$Var[\hat{p}]$	0.1871 (0.01)	0.1535 (0.01)	0.1589 (0.01)	0.0762 (0.01)	3 (0.02)
$MSE[\hat{a}]$	0.0199 (0.01)	0.2809 (0.01)	0.0045 (0.01)	0.0015 (0.01)	6.67×10^{-5} (0.02)
$MSE\left[\hat{b}\right]$	95 (0.06)	426 (0.08)	317 (0.05)	826 (0.03)	11 (0.07)
$MSE[\hat{p}]$	0.2486 (0.01)	0.2398 (0.01)	0.2635 (0.01)	0.2649 (0.01)	4 (0.02)

Table 3.39: Performance of the method of Appell moments (with $\alpha = 5$) for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$ for different sample size n_o.

3.5.4 Asymptotic Covariance Matrix of the Appell Moments Estimators

In this subsection, we illustrate the construction of the covariance matrix of the Appell moment estimators for a two-component exponential mixture, particularly for $\alpha = 3$. From the simulation results reported in previous section, we found poor estimation when $\alpha = 4$ and 5; therefore we only focus on $\alpha = 3$ in this section. Indeed, the covariance matrix of the Appell moment estimators for any larger α can be calculated in a similar procedure shown in the followings. By now, we are familiar with the construction of the covariance matrix of moment estimators, given by

$$\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right] \approx \boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1} \boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right] \left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{T}\right)^{-1}. \tag{3.160}$$

Hence, we need the Jacobian matrix of the moment estimator $\boldsymbol{D}\left[\boldsymbol{\Theta}\right]$ and the covariance matrix of the Appell moments $\boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right]$. The covariance of two Appell moments $\hat{\delta}_k$ and $\hat{\delta}_l$ is given by

$$Cov\left[\hat{\delta}_k, \hat{\delta}_l\right] = \frac{1}{n_o}\left(E\left[h_k\left(T\right) h_l\left(T\right)\right] - E\left[h_k\left(T\right)\right] E\left[h_l\left(T\right)\right]\right), \tag{3.161}$$

for $k = 0, 1, 2, 3$ and $l = 0, 1, 2, 3$, where $h_k\left(T\right)$'s are defined by (3.151) to (3.148). The followings are the elements of the covariance matrix of the Appell moments, $\boldsymbol{V}\left[\hat{\boldsymbol{\mu}}\right]$:

$$Var\left[\hat{\delta}_0\right] = \frac{1}{n_o}\left[\begin{array}{c} \left(\dfrac{9p\varpi^6}{32}\begin{bmatrix} 82 \\ -\dfrac{17}{1+4\varpi^2a^{-2}} \\ -\dfrac{18}{1+16\varpi^2a^{-2}} \\ +\dfrac{81}{1+36\varpi^2a^{-2}} \end{bmatrix} + \dfrac{9\left(1-p\right)\varpi^6}{32}\begin{bmatrix} 82 \\ -\dfrac{17}{1+4\varpi^2b^{-2}} \\ -\dfrac{18}{1+16\varpi^2b^{-2}} \\ +\dfrac{81}{1+36\varpi^2b^{-2}} \end{bmatrix}\right) \\ -\left(\dfrac{6p\varpi^3}{(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})} + \dfrac{6\left(1-p\right)\varpi^3}{(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)^2 \end{array}\right],$$

$$Var\left[\hat{\delta}_1\right] = \frac{1}{n_o}\left[\begin{array}{c} \left(\dfrac{p\varpi^4}{32n_o}\begin{bmatrix} 90 \\ -\dfrac{63}{1+4\varpi^2a^{-2}} \\ +\dfrac{54}{1+16\varpi^2a^{-2}} \\ -\dfrac{81}{1+36\varpi^2a^{-2}} \end{bmatrix} + \dfrac{\left(1-p\right)\varpi^4}{32n_o}\begin{bmatrix} 90 \\ -\dfrac{63}{1+4\varpi^2b^{-2}} \\ +\dfrac{54}{1+16\varpi^2b^{-2}} \\ -\dfrac{81}{1+36\varpi^2b^{-2}} \end{bmatrix}\right) \\ -\left(\dfrac{6p\varpi^3}{(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}a^{-1} + \dfrac{6\left(1-p\right)\varpi^3}{(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}b^{-1}\right)^2 \end{array}\right],$$

$$Var\left[\hat{\delta}_2\right] = \frac{1}{n_o}\left[\begin{array}{c}\left(\frac{p\varpi^2}{32}\begin{bmatrix}18\\-\frac{9}{1+4\varpi^2a^{-2}}\\-\frac{18}{1+16\varpi^2a^{-2}}\\+\frac{9}{1+36\varpi^2a^{-2}}\end{bmatrix}+\frac{(1-p)\varpi^2}{32}\begin{bmatrix}18\\-\frac{9}{1+4\varpi^2b^{-2}}\\-\frac{18}{1+16\varpi^2b^{-2}}\\+\frac{9}{1+36\varpi^2b^{-2}}\end{bmatrix}\right)\\-\left(\frac{6p\varpi^3}{(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}a^{-2}+\frac{6(1-p)\varpi^3}{(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}b^{-2}\right)^2\end{array}\right],$$

$$Var\left[\hat{\delta}_3\right] = \frac{1}{n_o}\left[\begin{array}{c}\left(\frac{p}{32}\begin{bmatrix}10\\-\frac{15}{1+4\varpi^2a^{-2}}\\+\frac{6}{1+16\varpi^2a^{-2}}\\-\frac{1}{1+36\varpi^2a^{-2}}\end{bmatrix}+\frac{(1-p)}{32}\begin{bmatrix}10\\-\frac{15}{1+4\varpi^2b^{-2}}\\+\frac{6}{1+16\varpi^2b^{-2}}\\-\frac{1}{1+36\varpi^2b^{-2}}\end{bmatrix}\right)\\-\left(\frac{6p\varpi^3}{(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}a^{-3}+\frac{6(1-p)\varpi^3}{(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}b^{-3}\right)^2\end{array}\right],$$

$$Cov\left[\hat{\delta}_0,\hat{\delta}_1\right] = \frac{1}{n_o}\left[\begin{array}{c}\left(\frac{3p\varpi^5}{32}\begin{bmatrix}\frac{42a\omega}{1+4\varpi^2a^{-2}}\\-\frac{144a\omega}{1+16\varpi^2a^{-2}}\\+\frac{486a\omega}{1+36\varpi^2a^{-2}}\end{bmatrix}+\frac{3(1-p)\varpi^5}{32}\begin{bmatrix}\frac{42a\omega}{1+4\varpi^2a^{-2}}\\-\frac{144a\omega}{1+16\varpi^2a^{-2}}\\+\frac{486a\omega}{1+36\varpi^2a^{-2}}\end{bmatrix}\right)\\-\left[\begin{array}{c}\left(\frac{6p\varpi^3}{(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\\\times\left(\frac{6p\varpi^3}{a(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\end{array}\right]\end{array}\right],$$

$$Cov\left[\hat{\delta}_0,\hat{\delta}_2\right] = \frac{1}{n_o}\left[\begin{array}{c}\left(-\frac{3p\varpi^4}{32}\begin{bmatrix}30\\-\frac{27}{1+4\varpi^2a^{-2}}\\-\frac{30}{1+16\varpi^2a^{-2}}\\+\frac{27}{1+36\varpi^2a^{-2}}\end{bmatrix}-\frac{3(1-p)\varpi^4}{32}\begin{bmatrix}30\\-\frac{27}{1+4\varpi^2a^{-2}}\\-\frac{30}{1+16\varpi^2a^{-2}}\\+\frac{27}{1+36\varpi^2a^{-2}}\end{bmatrix}\right)\\-\left[\begin{array}{c}\left(\frac{6p\varpi^3}{(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\\\times\left(\frac{6p\varpi^3}{a^2(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b^2(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\end{array}\right]\end{array}\right],$$

$$Cov\left[\hat{\delta}_0,\hat{\delta}_3\right]=\frac{1}{n_o}\left[\begin{array}{c}\left(-\frac{3p\varpi^3}{32}\left[\begin{array}{c}\frac{58a\omega}{1+4\varpi^2a^{-2}}\\-\frac{112a\omega}{1+16\varpi^2a^{-2}}\\+\frac{54a\omega}{1+36\varpi^2a^{-2}}\end{array}\right]-\frac{3(1-p)\varpi^3}{32}\left[\begin{array}{c}\frac{58a\omega}{1+4\varpi^2a^{-2}}\\-\frac{112a\omega}{1+16\varpi^2a^{-2}}\\+\frac{54a\omega}{1+36\varpi^2a^{-2}}\end{array}\right]\right)\\-\left[\begin{array}{c}\left(\frac{6p\varpi^3}{(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\\\times\left(\frac{6p\varpi^3}{a^3(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b^3(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\end{array}\right]\end{array}\right],$$

$$Cov\left[\hat{\delta}_1,\hat{\delta}_2\right]=\frac{1}{n_o}\left[\begin{array}{c}\left(\frac{-p\varpi^3}{32}\left[\begin{array}{c}\frac{-18a\omega}{1+4\varpi^2a^{-2}}\\-\frac{144a\omega}{1+16\varpi^2a^{-2}}\\+\frac{162a\omega}{1+36\varpi^2a^{-2}}\end{array}\right]-\frac{(1-p)\varpi^3}{32}\left[\begin{array}{c}\frac{-18a\omega}{1+4\varpi^2b^{-2}}\\-\frac{144a\omega}{1+16\varpi^2b^{-2}}\\+\frac{162a\omega}{1+36\varpi^2b^{-2}}\end{array}\right]\right)\\-\left[\begin{array}{c}\left(\frac{6p\varpi^3}{a(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\\\times\left(\frac{6p\varpi^3}{a^2(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b^2(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\end{array}\right]\end{array}\right],$$

$$Cov\left[\hat{\delta}_1,\hat{\delta}_3\right]=\frac{1}{n_o}\left[\begin{array}{c}\left(\frac{-p\varpi^2}{32}\left[\begin{array}{c}18-\frac{39}{1+4\varpi^2a^{-2}}\\+\frac{30}{1+16\varpi^2a^{-2}}\\-\frac{9}{1+36\varpi^2a^{-2}}\end{array}\right]-\frac{(1-p)\varpi^2}{32}\left[\begin{array}{c}18-\frac{39}{1+4\varpi^2b^{-2}}\\+\frac{30}{1+16\varpi^2b^{-2}}\\-\frac{9}{1+36\varpi^2b^{-2}}\end{array}\right]\right)\\-\left[\begin{array}{c}\left(\frac{6p\varpi^3}{a(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\\\times\left(\frac{6p\varpi^3}{a^3(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b^3(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\end{array}\right]\end{array}\right],$$

$$Cov\left[\hat{\delta}_2,\hat{\delta}_3\right]=\frac{1}{n_o}\left[\begin{array}{c}\left(\frac{p\varpi}{32}\left[\begin{array}{c}\frac{30a\omega}{1+4\varpi^2a^{-2}}\\-\frac{48a\omega}{1+16\varpi^2a^{-2}}\\+\frac{18a\omega}{1+36\varpi^2a^{-2}}\end{array}\right]+\frac{(1-p)\varpi}{32}\left[\begin{array}{c}\frac{30a\omega}{1+4\varpi^2b^{-2}}\\-\frac{48a\omega}{1+16\varpi^2b^{-2}}\\+\frac{18a\omega}{1+36\varpi^2b^{-2}}\end{array}\right]\right)\\-\left[\begin{array}{c}\left(\frac{6p\varpi^3}{a^2(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b^2(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\\\times\left(\frac{6p\varpi^3}{a^3(1+\varpi^2a^{-2})(1+9\varpi^2a^{-2})}+\frac{6(1-p)\varpi^3}{b^3(1+\varpi^2b^{-2})(1+9\varpi^2b^{-2})}\right)\end{array}\right]\end{array}\right].$$

When $\alpha = 3$, we estimate the three parameters of a two-component exponential mixture distribution by equating four Appell sample moments to their theoretical counterparts. The

Jacobian matrix $\boldsymbol{D}[\boldsymbol{\Theta}]$ is a 4×3 matrix with entries $d_{ij} = \dfrac{\partial \delta_i}{\partial \Theta_j}$ from (3.162) to (3.164) where $i = 0, 1, 2, 3$, where

$$\frac{\partial \delta_i}{\partial a} = \frac{p\varpi^5}{a^{3+i}} \left[\begin{array}{c} \dfrac{108}{(1+\varpi^2 a^{-2})(1+9\varpi^2 a^{-2})^2} \\ + \dfrac{12}{(1+\varpi^2 a^{-2})^2(1+9\varpi^2 a^{-2})} \end{array} \right] - \frac{6ip\varpi^3}{a^{i+1}(1+\varpi^2 a^{-2})(1+9\varpi^2 a^{-2})}, \tag{3.162}$$

$$\frac{\partial \delta_i}{\partial b} = \frac{q\varpi^5}{b^{3+i}} \left[\begin{array}{c} \dfrac{108}{(1+\varpi^2 b^{-2})(1+9\varpi^2 b^{-2})^2} \\ + \dfrac{12}{(1+\varpi^2 b^{-2})^2(1+9\varpi b^{-2})} \end{array} \right] - \frac{6iq\varpi^3}{b^{i+1}(1+\varpi^2 b^{-2})(1+9\varpi^2 b^{-2})}, \tag{3.163}$$

$$\frac{\partial \delta_i}{\partial p} = \frac{6\varpi^3}{a^i(1+\varpi^2 a^{-2})(1+9\varpi^2 a^{-2})} - \frac{6\varpi^3}{b^i(1+\varpi^2 b^{-2})(1+9\varpi^2 b^{-2})}. \tag{3.164}$$

So, the $\boldsymbol{D}[\boldsymbol{\Theta}]$ in (3.160) is not a square matrix and hence is not invertible. To solve (3.160) we consider the two approaches used in previous section to solve $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ of the attenuated moment estimator: First, we add an extra column to $\boldsymbol{D}[\boldsymbol{\Theta}]$ with $\dfrac{\partial \delta_i}{\partial q}$, given by

$$\frac{\partial \delta_i}{\partial q} = \frac{6\varpi^3}{b^i(1+\varpi^2 b^{-2})(1+9\varpi^2 b^{-2})}, \tag{3.165}$$

and amend $\dfrac{\partial \delta_i}{\partial p}$ in (3.164) so that it becomes

$$\frac{\partial \delta_i}{\partial p} = \frac{6\varpi^3}{a^i(1+\varpi^2 a^{-2})(1+9\varpi^2 a^{-2})}, \tag{3.166}$$

and we call this approach the Q-version of $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$.

Secondly, we find the generalised inverse of $\boldsymbol{D}[\boldsymbol{\Theta}]$ given by

$$\boldsymbol{D}[\boldsymbol{\Theta}]^{-1} = \left(\boldsymbol{D}[\boldsymbol{\Theta}]^T \boldsymbol{D}[\boldsymbol{\Theta}]\right)^{-1} \boldsymbol{D}[\boldsymbol{\Theta}]^T \tag{3.167}$$

and

$$\left(\boldsymbol{D}[\boldsymbol{\Theta}]^T\right)^{-1} = \boldsymbol{D}[\boldsymbol{\Theta}] \left(\boldsymbol{D}[\boldsymbol{\Theta}]^T \boldsymbol{D}[\boldsymbol{\Theta}]\right)^{-1} \tag{3.168}$$

to make $\boldsymbol{D}[\boldsymbol{\Theta}]^{-1}$ a 3×4 matrix and $\left(\boldsymbol{D}[\boldsymbol{\Theta}]^T\right)^{-1}$ a 4×3 matrix. By doing this, $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ remains as a 3×3 matrix. We name this approach the GI-version of $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$.

We calculate these two versions' theoretical variances of the estimators and compare them to check if they agree with each other. Then, we move on to find conformity between simulated and theoretical results. Previously, we have seen that the Q-version has a better agreement to the simulation results. It is worth to investigate if the Q-version also outperforms the GI-version in this case.

r	Theoretical $Var[\hat{a}]_Q$	Theoretical $Var[\hat{a}]_{GI}$	Practical $Var[\hat{a}]$
2	0.0003 ($\omega = 0.0250$)	0.0003 ($\omega = 0.0248$)	0.0003 ($\omega = 0.01$)
5	6.86×10^{-5} ($\omega = 0.0375$)	6.88×10^{-5} ($\omega = 0.0371$)	7.01×10^{-5} ($\omega = 0.04$)
10	5.02×10^{-5} ($\omega = 0.0400$)	5.03×10^{-5} ($\omega = 0.0400$)	5.05×10^{-5} ($\omega = 0.04$)

Table 3.40: Theoretical and simulated minimum variance of Appell moment estimator (with $\alpha = 3$) $\hat{a}$ given by the optimal ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical $Var\left[\hat{b}\right]_Q$	Theoretical $Var\left[\hat{b}\right]_{GI}$	Practical $Var\left[\hat{b}\right]$
2	0.0065 ($\omega = 0.0273$)	0.0065 ($\omega = 0.0290$)	21 ($\omega = 0.05$)
5	0.0105 ($\omega = 0.0923$)	0.0136 ($\omega = 0.1826$)	0.0173 ($\omega = 0.08$)
10	0.0310 ($\omega = 0.2001$)	0.0284 ($\omega = 0.3127$)	0.0444 ($\omega = 0.20$)

Table 3.41: Theoretical and simulated minimum variance of Appell moment estimator (with $\alpha = 3$) $\hat{b}$ given by the optimal ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

3.5.5 Optimal ω

In this subsection, we check the accuracy of the two versions of approximated theoretical variance of the estimator by comparing them to the simulation results. We first use the Mathematica function "FindMinimum" to obtain the minimum $Var[\hat{a}]$, $Var\left[\hat{b}\right]$, $Var[\hat{p}]$ and their ω-counterparts for both the Q-version and the GI-version. Having done that, we use the simulation results in Section 3.5.3 to investigate the conformity between simulated and theoretical results, presented in Tables 3.40, 3.41 and 3.42.

From Table 3.40, we observe that both $Var[\hat{a}]_Q$ and $Var[\hat{a}]_{GI}$ have good agreements with each other in terms of both the values and the optimal ω. Excitingly, for all three degrees of separation considered, the practical $Var[\hat{a}]$ and the best ω are both close to the theoretical values, and the agreement is better for larger r. However, for b (see Table 3.41), the GI-version has larger optimal ω's; whereas the Q-version has a better agreement to the practical $Var\left[\hat{b}\right]$. Similarly, for p (see Table 3.42), the Q-version reflects a better story of the practical $Var[\hat{p}]$.

Since the Q-Version has a somewhat different agreement with the GI-Version regarding the minimum variance of Appell Sequences estimator of b and p, we estimate another

r	Theoretical $Var[p]_Q$	Theoretical $Var[p]_{GI}$	Practical $Var[\hat{p}]$
2	0.1345 ($\omega = 0.0273$)	0.1345 ($\omega = 0.0270$)	0.0765 ($\omega = 0.02$)
5	0.0047 ($\omega = 0.0826$)	0.0044 ($\omega = 0.1159$)	0.0049 ($\omega = 0.06$)
10	0.0016 ($\omega = 0.1074$)	0.0015 ($\omega = 0.2045$)	0.0016 ($\omega = 0.10$)

Table 3.42: Theoretical and simulated minimum variance of Appell moment estimator (with $\alpha = 3$) $\hat{p}$ given by the optimal ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Q Version	GI Version
2	$\omega = 0.0273$	$\omega = 0.0290$
	Theoretical $Var\left[\hat{b}\right]_Q = 0.0065$ Practical $Var\left[\hat{b}\right] = 6$	Theoretical $Var\left[\hat{b}\right]_{GI} = 0.0065$ Practical $Var\left[\hat{b}\right] = 7$
5	$\omega = 0.0923$	$\omega = 0.1826$
	Theoretical $Var\left[\hat{b}\right]_Q = 0.0105$ Practical $Var\left[\hat{b}\right] = 0.0164$	Theoretical $Var\left[\hat{b}\right]_{GI} = 0.0136$ Practical $Var\left[\hat{b}\right] = 0.1159$
10	$\omega = 0.2001$	$\omega = 0.3127$
	Theoretical $Var\left[\hat{b}\right]_Q = 0.0310$ Practical $Var\left[\hat{b}\right] = 0.0429$	Theoretical $Var\left[\hat{b}\right]_{GI} = 0.0284$ Practical $Var\left[\hat{b}\right] = 0.2598$

Table 3.43: Checking the accuracy of two versions of theoretical variance of Appell moment estimator $\hat{b}$ (with $\alpha = 3$) for a mixture of two exponential distributions with $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

10000 simulated samples with the suggested best ω and find the practical variances of the estimators. In Table 3.43, we can see that for large r (e.g. $r = 5$ and $r = 10$), the optimal ω for estimating b suggested by the GI-Version is larger than the one suggested by the Q-Version. In practice, the simulation results show that the Q-Version provides a more realistic picture because the practical $Var\left[\hat{b}\right]$'s have values close to the theoretical values. Conversely, the GI-Version fails to provide a reliable suggestion because we do not get the ideal small $Var\left[\hat{b}\right]$ using the suggested ω. The GI-Version actually under-estimates $Var[b]$ and over-estimates the optimal ω.

Similarly, for p, the Q-Version tells a better story of the practical estimation than the GI-Version, as seen in Table 3.44. The GI-Version over-estimates the optimal ω and under-estimates the variance of the estimator for large r. Notably. for $r = 10$, we used $\omega = 0.1074$ to estimate 10000 simulated samples and the resulting $Var[\hat{p}]$ is exactly the same as the one estimated by the Q-Version. This $Var[\hat{p}]$ is indeed smaller than the minimum $Var[\hat{p}]$ we get when using $\omega = 0.1$.

r	Q Version	GI Version
2	$\omega = 0.0273$	$\omega = 0.0270$
	Theoretical $Var[\hat{p}]_Q = 0.1345$ Practical $Var[\hat{p}] = 0.0760$	Theoretical $Var[\hat{p}]_{GI} = 0.1345$ Practical $Var[\hat{p}] = 0.0776$
5	$\omega = 0.0826$	$\omega = 0.1159$
	Theoretical $Var[\hat{p}]_Q = 0.0047$ Practical $Var[\hat{p}] = 0.0052$	Theoretical $Var[\hat{p}]_{GI} = 0.0044$ Practical $Var[\hat{p}] = 0.0058$
10	$\omega = 0.1074$	$\omega = 0.2045$
	Theoretical $Var[\hat{p}]_Q = 0.0016$ Practical $Var[\hat{p}] = 0.0016$	Theoretical $Var[\hat{p}]_{GI} = 0.0015$ Practical $Var[\hat{p}] = 0.0139$

Table 3.44: Checking the accuracy of two versions of theoretical variance of Appell moment estimator $\hat{p}$ (with $\alpha = 3$) for a mixture of two exponential distributions with $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

Once again, the Q-Version provides a more precise approximation to $Var\left[\hat{\Theta}\right]$. Hence, we should use the optimal ω suggested by the Q-version to estimate the parameters of a mixture of two exponential distributions. Tables 3.45, 3.46 and 3.47 compares the Q-Version theoretical $Var[\hat{a}]_Q$, $Var\left[\hat{b}\right]_Q$ and $Var[\hat{p}]_Q$ with the practical values for nine degrees of separation, r ranging from 2 to 10. For each r, 10000 data sets each with 1000 observations were simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.1r$ and $p = 0.6$. We then estimate from each data set with the nineteen values of ω considered in previous subsection and we find the minimum variance for each case. Having done that, we estimate using the data sets with the theoretical best ω suggested by the Q-Version and compare the resulted variance with the minimum variance (out of the 19 sets) we obtained earlier. The minimum variance and its counterpart ω are shown on these three tables. In Table 3.45, the practical results have a good agreement with the theoretical values except from the ω for $r = 2$. Our simulation results show that the best ω for estimating a when $r = 2$ is 0.01 instead of the suggested 0.0250. As shown in Table 3.46, the best ω suggested by the Q-Version does in practice give us the estimates of b with lowest variation. In fact, the Q-Version under-estimates $Var\left[\hat{b}\right]$; from the simulation results, we observe slightly larger practical $Var\left[\hat{b}\right]$'s compared to the theoretical values. It is worth noting that $Var\left[\hat{b}\right]$ is getting larger when r increases from $r = 3$. From Table 3.47, we observe that, except from $r = 5$, the ω's suggested by the Q-Version provide the minimum $Var[\hat{p}]$'s for each r in our simulation experiments. $Var[\hat{p}]$ decreases when the separation between the components gets larger. Our study is based on mixtures of exponentials with fixed $a = 0.1$, the optimal values of ω changes proportionally with a. For instance, when $r = 10$, the best ω for estimating p is 0.1074 if the true parameter a is 0.1; if a is 1 and b is 10, the optimal ω is then increased to 1.0736.

Figure 3.23 shows the plots of $Var\left[\hat{\Theta}\right]_Q$ with respect to ω for different b ranging from 0.2 to 1. We observe that the plots of the theoretical $Var\left[\hat{\Theta}\right]_Q$ are symmetrical at $\omega = 0$.

r	Theoretical Optimal ω	Practical Optimal ω	Theoretical $Var\left[\hat{a}\right]_Q$	Practical $Var\left[\hat{a}\right]$
2	0.0250	0.0100	0.0003	0.0003
3	0.0323	0.0323	0.0001	0.0001
4	0.0357	0.0357	8.31×10^{-5}	8.28×10^{-5}
5	0.0375	0.0375	6.86×10^{-5}	6.96×10^{-5}
6	0.0385	0.0385	6.12×10^{-5}	6.02×10^{-5}
7	0.0391	0.0391	5.68×10^{-5}	5.74×10^{-5}
8	0.0395	0.0395	5.39×10^{-5}	5.54×10^{-5}
9	0.0398	0.0398	5.18×10^{-5}	5.29×10^{-5}
10	0.0400	0.0400	5.02×10^{-5}	4.94×10^{-5}

Table 3.45: Theoretical and simulated variance of Appell moment estimator (with $\alpha = 3$) $\hat{a}$ and counterpart-ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical Optimal ω	Practical Optimal ω	Theoretical $Var\left[\hat{b}\right]$	Practical $Var\left[\hat{b}\right]$
2	0.0298	0.0298	0.0065	6
3	0.0495	0.0495	0.0061	0.0490
4	0.0705	0.0705	0.0079	0.1222
5	0.0923	0.0923	0.0105	0.0164
6	0.1141	0.1141	0.0136	0.0198
7	0.1358	0.1358	0.0173	0.0253
8	0.1574	0.1574	0.0214	0.0291
9	0.1788	0.1788	0.0259	0.0348
10	0.2001	0.2001	0.0310	0.0429

Table 3.46: Theoretical and simulated variance of Appell moment estimator (with $\alpha = 3$) $\hat{b}$ and counterpart-ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical Optimal ω	Practical Optimal ω	Theoretical $Var\left[\hat{p}\right]$	Practical $Var\left[\hat{p}\right]$
2	0.0273	0.0273	0.1345	0.0760
3	0.0409	0.0409	0.0191	0.0176
4	0.0552	0.0552	0.0081	0.0079
5	0.0826	0.0600	0.0047	0.0049
6	0.1214	0.1214	0.0030	0.0034
7	0.1173	0.1173	0.0022	0.0025
8	0.1131	0.1131	0.0019	0.0020
9	0.1099	0.1099	0.0017	0.0017
10	0.1074	0.1074	0.0016	0.0016

Table 3.47: Theoretical and simulated variance of Appell moment estimator (with $\alpha = 3$) $\hat{p}$ and counterpart-ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

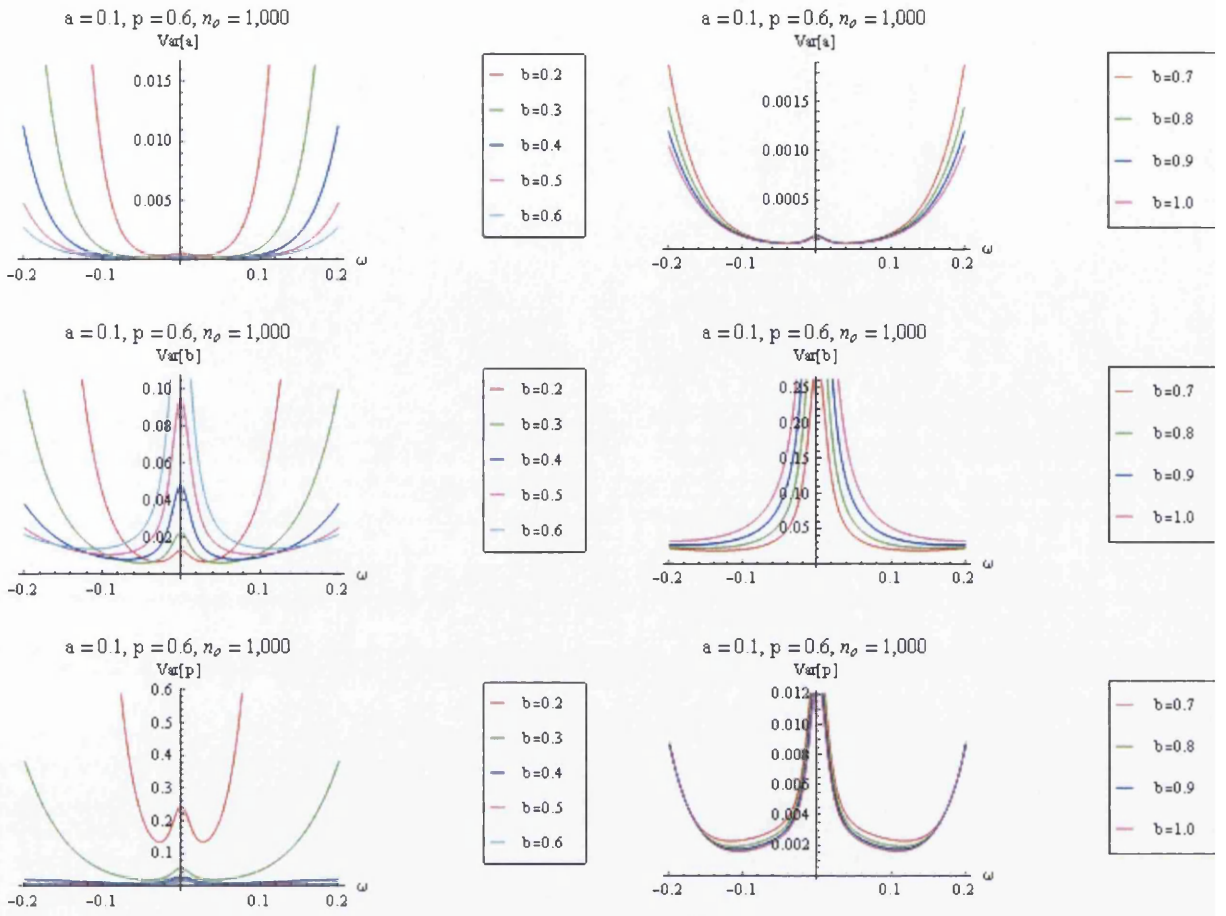


Figure 3.23: Plots of theoretical $Var\left[\hat{\Theta}\right]_Q$ versus ω for varying r.

This means that we could also use negative ω to estimate Θ. The best ω's for estimating Θ have larger absolute values for mixed exponential distributions with larger separation between the two components.

Our simulation experiments did not make use of negative ω, so we carried out another simulation experiment in which we use the suggested optimal negative ω and compare the practical variances of estimators with the theoretical ones; the results are presented in Tables 3.48, 3.49 and 3.50. The simulation results confirm that negative ω's are equally good as the positive ω. The larger is the separation between the two components, the better is the agreement between the simulated values and the theoretical values.

3.5.6 Discussion

To summarise, using the Appell Sequences estimator with $\alpha = 3$, users are suggested to take $\omega = 0.01$ or 0.02 to estimate all parameters (a, b and p) of mixture exponential distributions when the sample size is small ($n_o \leq 50$). Like the other methods, the estimation of b is

r	ω	Theoretical $Var\left[\hat{a}\right]_Q$	Practical $Var\left[\hat{a}\right]$
2	−0.0250	0.0003	0.0004
5	−0.0375	6.86×10^{-5}	7.04×10^{-5}
10	−0.0400	5.02×10^{-5}	5.03×10^{-5}

Table 3.48: Theoretical and simulated variance of Appell moment estimator (with $\alpha = 3$) $\hat{a}$ given by negative ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	ω	Theoretical $Var\left[\hat{b}\right]_Q$	Practical $Var\left[\hat{b}\right]$
2	−0.0273	0.0065	6
5	−0.0923	0.0105	0.0165
10	−0.2001	0.0310	0.0450

Table 3.49: Theoretical and simulated variance of Appell moment estimator (with $\alpha = 3$) $\hat{b}$ given by negative ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	ω	Theoretical $Var\left[\hat{p}\right]_Q$	Practical $Var\left[\hat{p}\right]_Q$
2	−0.0273	0.1345	0.0770
5	−0.0826	0.0047	0.0050
10	−0.1074	0.0016	0.0016

Table 3.50: Theoretical and simulated variance of Appell moment estimator (with $\alpha = 3$) $\hat{p}$ given by negative ω for a mixture of two exponential distributions with various r and fixed $a = 0.1$, $b = 0.1r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications

not satisfying for samples of small sizes. When samples are large enough, say $n_o = 1000$, we should use the asymptotic theoretical optimal ω suggested in Section 3.5.5 to achieve estimates with high precision.

Like any moment based method, we should note that there is no guarantee that the roots of the determinantal equation in (3.130) are positive real numbers. Hence, this method might provide estimates with negative values or in complex forms. As mentioned before, $\mathbf{V}_\alpha^T \mathbf{\Lambda} \mathbf{V}_\alpha$ in (3.140) is likely to be a singular matrix and hence we might not get an appropriate estimate of the mixing weight p.

In conclusion, the method based on Appell sequences should, in many cases, be able to provide reasonable parameter estimates, provided that the samples are large enough, especially in the case when component distributions are not well separated.

3.6 Method Using Order Statistics

3.6.1 Introduction

In this section we study a method using order statistics, devised by Jalali (2007), for the parameter estimation of a two-component exponential distribution. The theoretical exposition in this subsection follows his paper. Let X and Y be two continuous random variables with CDFs F_X and F_Y, and the PDFs f_X and f_Y, then

$$\Pr[X \leq Y] = \int_{-\infty}^{\infty} f_Y(y)\, dy \int_{-\infty}^{y} f_X(x)\, dx = \int_{-\infty}^{\infty} f_Y(y)\, F_X(y)\, dy.$$

Alternatively,

$$\Pr[X \leq Y] = \int_{-\infty}^{\infty} f_X(x)\, dx \int_{x}^{\infty} f_Y(y)\, dy = \int_{-\infty}^{\infty} f_X(x)\, S_Y(x)\, dx,$$

where $S_Y(y) = 1 - F_Y(y)$ is Y's survival function.

Now let X_i's be a sample of size $\varkappa$ from the exponential distribution with mean a^{-1}, Y_i a sample of size η from the exponential distribution with mean b^{-1},

$$\begin{aligned} X &= \min[X_i], \\ Y &= \min[Y_i], \end{aligned}$$

and

$$Z = \min[X, Y].$$

Then

$$\Pr[X \leq Y] = \varkappa a \int_0^{\infty} \exp(-\varkappa a x) \exp(-\eta b x)\, dx = \frac{\varkappa a}{\varkappa a + \eta b}.$$

This is also the probability of the minimum element of the union of the two samples coming

from the first sample. Z is an exponential distribution with mean $\frac{1}{\varkappa a+\eta b}$, hence,

$$E[Z]=\frac{1}{\varkappa a+\eta b}.$$

But the conditional expectations of Z given that $X \leq Y$ is

$$E[Z|X \leq Y]=\frac{\varkappa a \int_0^\infty x \exp(-\varkappa a x)\exp(-\eta b x)\,dx}{\frac{\varkappa a}{\varkappa a+\eta b}}=\frac{1}{\varkappa a+\eta b}.$$

This means that

$$E[Z|X > Y]=\frac{1}{\varkappa a+\eta b}.$$

So the expectation of Z is the same, whether Z belongs to the first or the second sample. In general, this result may not be true but it holds for exponential distributions because of the memoryless property.

Now let M_i be a sample of size n_s of a mixture of the two exponential distributions with means, respectively, of a^{-1} and b^{-1}, and mixing weights p and q. Let M be the minimum of this sample, it is easy to see that

$$E[M]=\sum_{\varkappa=0}^{n_s}\frac{\binom{n_s}{\varkappa}p^{\varkappa}q^{n_s-\varkappa}}{\varkappa a+(n_s-\varkappa)b}.$$

For $n_s=2$ and $n_s=3$ these expectations are, respectively, $\frac{p^2}{2a}+\frac{2pq}{a+b}+\frac{q^2}{2b}$ and $\frac{p^3}{3a}+\frac{3p^2q}{2a+b}+\frac{3pq^2}{a+2b}+\frac{q^3}{3b}$.

Estimating a Mixture of Two Exponential Distributions with Known p

Now suppose we have an "efficient" method of finding the sample means of the minimum of a pair and a triplet of the elements of our mixed distribution. Let these be equal to 3ϕ and 2τ. We suppose also we have the usual sample mean of such a mixed sample which we denote by 6η. Suppose first that p (and thus $q=1-p$) is given. Then an estimate of a and b can be obtained by solving the following equations for a and b:

$$\begin{aligned}\frac{p}{a}+\frac{q}{b} &= 6\eta, \\ \frac{p^2}{2a}+\frac{2pq}{a+b}+\frac{q^2}{2b} &= 3\phi.\end{aligned} \tag{3.169}$$

The roots in (3.170) give us estimates of a and b. We set $x = \frac{1}{6a}$ and $y = \frac{1}{6b}$, and the two equations in (3.169) are reduced to

$$\begin{aligned} px + qy &= \eta, \\ (px + qy)^2 + xy &= \phi(x + y). \end{aligned} \tag{3.170}$$

Eliminating y from (3.170) gives us the following quadratic equation in terms of x:

$$px^2 - x(\eta + \phi(p - q) + \eta(\phi - q\eta)) = 0. \tag{3.171}$$

Before proceeding further, we shall prove two inequalities in the form of the following lemma.

Lemma 6 *If there are positive a and b as above, then $\eta \geq \phi$. If further, $p \geq q$, then $\phi > q\eta$.*

Proof. For the first inequality it is sufficient to note that

$$\eta - \phi = pq\left(\frac{1}{6a} - \frac{2}{3(a+b)} + \frac{1}{6b}\right) = \frac{pq(a-b)^2}{6ab(a+b)} \geq 0.$$

For the second inequality we note that

$$\phi - q\eta = \frac{2pq}{6(a+b)} + \frac{p(p-q)}{6a} > 0$$

if $p \geq q$. ■

The roots of (3.171) are

$$x = \frac{1}{2p}\left[\eta + \phi(p - q) \pm \sqrt{(\eta - \phi)[2\eta(1 - 2z) - 4\phi z]}\right],$$

where $z = \frac{1 - 4pq}{4} \geq 0$. Then,

$$y = \frac{1}{2q}\left[\eta + \phi(q - p) \mp \sqrt{(\eta - \phi)[2\eta(1 - 2z) - 4\phi z]}\right]$$

which are the roots of equation

$$qy^2 - y(\eta + \phi(q - p)) + \eta(\phi - p\eta) = 0.$$

From the lemma it follows that at least one of these two equations has two positive roots. Hence we can find at least one pair of positive solutions for x and y. Then the corresponding a and b can be obtained, but x and y are more basic.

Estimating a Mixture of Two Exponential Distributions with Unknown p

Now, suppose the probability weights are also unknown. We need three equations to estimate all parameters. We choose the following three:

$$\begin{aligned} \frac{p}{a}+\frac{q}{b} &= 6\eta, \\ \frac{p^2}{2a}+\frac{2pq}{a+b}+\frac{q^2}{2b} &= 3\phi, \\ \frac{p^3}{3a}+3pq\left(\frac{p}{2a+b}+\frac{q}{a+2b}\right)+\frac{q^3}{3b} &= 2\tau. \end{aligned} \tag{3.172}$$

Here, 6η is the sample mean, 3ϕ is the mean of minimum of sample pairs, and 2τ is the mean of the minimum of sample triplets.

Later in this subsection, we shall show that the inequalities $\eta \geq \phi \geq \tau$ should hold for our system of equations to have non-negative solutions for a, b and p, q $(=1-p)$. (In fact the result is much more general than this.)

In order to solve equations (3.172), we let $x=\frac{1}{6a}$ and $y=\frac{1}{6b}$, which are the means of our exponential distributions. The system (3.172), then, reduces to the following

$$\begin{aligned} px+qy &= \eta, \\ \left(p^2x+q^2y\right)(x+y)+4pqxy &= \phi(x+y), \\ \left(p^3x+q^3y\right)(x+2y)(2x+y)+9pqxy\left(p(2x+y)+q(x+2y)\right) &= \tau(x+2y)(2x+y). \end{aligned} \tag{3.173}$$

If we eliminate p (and q) between the first two equations and set $\sigma=x+y$ and $\pi=xy$, we shall have the following equation:

$$\pi=\phi\sigma-\eta^2. \tag{3.174}$$

Eliminating p between the second and the third equations in (3.173) results in the following equations:

$$2\eta^3+3\eta\pi-\left(2\sigma^2+\pi\right)\tau+2\pi\sigma=0. \tag{3.175}$$

Replacing π (given by (3.174)) in (3.175) results in the following quadratic equation for σ:

$$f(\sigma)=2\eta^2+\left(3\eta\phi-2\eta^2-\phi\tau\right)\sigma-\eta^2(\eta-\tau)=0. \tag{3.176}$$

If $\eta>\phi>\tau$, then (3.176) has one and only one positive root, as

$$\frac{-\eta^2(\eta-\tau)}{\phi-\tau}<0.$$

Using this positive root σ we can obtain π from the (3.174); π is positive if and only if $\eta\tau>\phi^2$. The following establishes more clearly the necessity of the latter condition.

Lemma 7 *If our system of equations has positive solution for p and q and positive and*

distinct solutions for x and y, then it is necessary that $\eta\tau - \phi^2 > 0$.

Proof. By elementary manipulation one can show that

$$\eta\tau - \phi^2 = \frac{pqxy\,(x-y)^2\left[2\,(x+y)\left(p^2\,(2x+y)+q^2\,(x+2y)\right)+pq\left(5x^2+14xy+5y^2\right)\right]}{(x+2y)\,(2x+y)\,(x+y)^2}.$$

This is of course greater than zero, p and q are positive, x and y are positive and distinct. ■

The three conditions $\eta > \phi > \tau$ and $\eta\tau - \phi^2 > 0$ as shown are necessary for the existence of p, q, x and y for a proper mixture of exponentials. We saw that these conditions are also sufficient for the existence of positive σ and π. Now to obtain x and y we need to solve the quadratic equation

$$u^2 - \sigma u + \pi = 0. \tag{3.177}$$

This equation has two distinct positive roots if and only if its discriminant $\sigma^2 - 4\pi$ is positive. But from the aforesaid equation this can be written as

$$\sigma^2 - 4\phi\sigma + 4\eta^2 = (\sigma - 2\phi)^2 + 4\left(\eta^2 - \phi^2\right), \tag{3.178}$$

which is positive under our conditions.

The two roots of the last quadratic equation (3.177) are of course x and y. Now it only remain to find p and q. Clearly from the first equation in our system of three equations,

$$p = \frac{\eta - y}{x - y} \tag{3.179}$$

and

$$q = \frac{\eta - x}{y - x}. \tag{3.180}$$

But we know the mean η should be between the x and y, so p and q should be both positive. As we have $p + q = 1$, they are both also less than 1, and therefore proper probabilities. We now sum up what we have proved:

Theorem 8 *For our system of equations to have positive and distinct solutions for x and y and positive solutions for p and q, it is necessary and sufficient that $\eta > \phi > \tau$ and $\eta\tau - \phi^2 > 0$.*

Practical Estimation We shall now illustrate how we obtain estimates of 3ϕ and 2τ when given a raw sample $t_1, ..., t_{n_o}$. We first sort the data in order and let $t_{(i)}$ denote the i^{th} order statistic of the sample. We then substitute 3ϕ by

$$\hat{z}_2 = \frac{S_2}{\binom{n_o}{2}}, \tag{3.181}$$

where

$$S_2 = \sum_{i=1}^{n_o} (n_o - i)\, t_{(i)}, \tag{3.182}$$

and 2τ by

$$\hat{z}_3 = \frac{S_3}{\binom{n_o}{2}}, \tag{3.183}$$

where

$$S_3 = \sum_{i=1}^{n_o} \binom{n_o - i}{2} t_{(i)}. \tag{3.184}$$

In order to find (3.181) and (3.183), we need to re-sample the data, first into pairs of elements for $\hat{z}_2$ and then into triplets of elements for $\hat{z}_3$. For example, if we have a sample consisting of five observations, we first arrange them in order:

$$t_{(1)}, t_{(2)}, t_{(3)}, t_{(4)}, t_{(5)}$$

and re-sample them into sub-samples where each sub-sample consists of two elements. We first compare the smallest element $t_{(1)}$ with all other data:

$$\begin{matrix} (t_{(1)}, t_{(2)}) \\ (t_{(1)}, t_{(3)}) \\ (t_{(1)}, t_{(4)}) \\ (t_{(1)}, t_{(5)}) \end{matrix}$$

Obviously, for each pair of element, the minimum is $t_{(1)}$; therefore, the sum of minimum so far is $4t_{(1)}$. Comparing $t_{(2)}$ with the others, we know $t_{(2)}$ is the minimum in the following pairs:

$$\begin{matrix} (t_{(2)}, t_{(3)}) \\ (t_{(2)}, t_{(4)}) \\ (t_{(2)}, t_{(5)}) \end{matrix}$$

so sum of minimum is $3t_{(2)}$. Clearly, by comparing $t_{(3)}$ with the others, it is the minimum in the following pairs:

$$\begin{matrix} (t_{(3)}, t_{(4)}) \\ (t_{(3)}, t_{(5)}) \end{matrix}$$

and $t_{(4)}$ will only be minimum in the pair $(t_{(4)}, t_{(5)})$. Therefore, for the whole sample, we know that the sum of minimum in all sub-samples is

$$S_2 = 4t_{(1)} + 3t_{(2)} + 2t_{(3)} + t_{(4)}. \tag{3.185}$$

It is hence easy to see that (3.185) can simply be estimated with (3.182). To find $\hat{z}_2$, we re-sample the observations into ten $(= \begin{pmatrix} 5 \\ 3 \end{pmatrix})$ sub-samples, each with three elements, as follows:

$$\begin{array}{lll} (t_{(1)}, t_{(2)}, t_{(3)}) & (t_{(2)}, t_{(3)}, t_{(4)}) & (t_{(3)}, t_{(4)}, t_{(5)}) \\ (t_{(1)}, t_{(2)}, t_{(4)}) & (t_{(2)}, t_{(3)}, t_{(5)}) & \\ (t_{(1)}, t_{(2)}, t_{(5)}) & (t_{(2)}, t_{(3)}, t_{(6)}) & \\ (t_{(1)}, t_{(3)}, t_{(4)}) & & \\ (t_{(1)}, t_{(3)}, t_{(5)}) & & \\ (t_{(1)}, t_{(4)}, t_{(5)}) & & \end{array}$$

It is obvious that $t_{(1)}$ is the minimum in each triplet in the first column, $t_{(2)}$ is the minimum in each triplet in the second column and $t_{(3)}$ is the minimum in the triplet in the last column. Therefore, in this case

$$S_3 = 6t_{(1)} + 3t_{(2)} + t_{(3)},$$

obviously this can be found easily from (3.184).

Having found $\hat{z}_2$ and $\hat{z}_3$ according to (3.181) and (3.183), we estimate η, ϕ and τ from

$$\begin{aligned} \hat{\eta} &= \frac{\sum_{i=1}^{n_o} t_{(i)}}{6n_o}, \\ \hat{\phi} &= \frac{\hat{z}_2}{3}, \\ \hat{\tau} &= \frac{\hat{z}_3}{6}, \end{aligned}$$

and substitute them into (3.176) to solve for $\hat{\sigma}$, which is the positive root of $f(\sigma)$. After that, $\hat{\pi}$ can be found by substituting $\hat{\sigma}$, $\hat{\eta}$ and $\hat{\phi}$ into (3.174). Since both $\hat{\sigma}$ and $\hat{\pi}$ are known now, we substitute them into (3.177) to find the roots $\hat{x}$ and $\hat{y}$. Finally, the desired estimates of a, b and p can be obtained from $\frac{1}{6\hat{x}}$, $\frac{1}{6\hat{y}}$ and (3.179) respectively.

3.6.2 Simulation Results

For illustration, a simulation experiment was carried out to examine the performance of the method using order statistics. Again, we study exponential mixture distribution with different separation between the two populations, ranging from small ($r = 2$) over medium ($r = 5$) to large ($r = 10$). For each r, we generated 10000 data sets of size n_o arising from a mixture of two exponential distributions with $a = 0.1$, $b = 0.1r$ and $p = 0.6$. Like before, we consider various sample size $n_o = (10, 15, 20, 50, 1000)$. Tables 3.51 to 3.53 present the average of the parameter estimates and their associated measures of error. When the number of observations in a data set increases, the averages of the 10000 parameter estimates approach to the true values.

Overall, the estimation of a and p are satisfactory, especially when n_o increases. Remarkably, p is well estimated even when the number of observations is scarce, especially

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1168	0.1116	0.1101	0.1051	0.0931
$E[\hat{b}]$	−0.3762	0.1683	0.1676	0.2991	0.2231
$E[\hat{p}]$	0.8156	0.7998	0.7955	0.7638	0.5653
$(\bar{\hat{a}} - a)^2$	0.0003	0.0001	0.0001	2.63×10^{-5}	4.72×10^{-5}
$(\bar{\hat{b}} - b)^2$	0.3321	0.0010	0.0011	0.0098	0.0005
$(\bar{\hat{p}} - p)^2$	0.0465	0.0399	0.0382	0.0268	0.0012
$Var[\hat{a}]$	0.0045	0.0030	0.0025	0.0015	0.0007
$Var[\hat{b}]$	346	235	153	260	4
$Var[\hat{p}]$	0.2381	0.2706	0.1914	0.1901	0.0908
$MSE[\hat{a}]$	0.0048	0.0032	0.0026	0.0016	0.0007
$MSE[\hat{b}]$	346	235	153	260	4
$MSE[\hat{p}]$	0.2845	0.3105	0.2296	0.2170	0.0920

Table 3.51: Performance of the method using order statistics for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$ for different sample size n_o.

when the separation between the components is large (see Table 3.52 and 3.53). In other words, this estimator is good at "spotting" a mixture distribution by providing an accurate estimate of the mixing weight, even when the number of data is limited. Like the estimators discussed earlier, the estimation of b is implausible for samples of small sizes. However, the variance of $\hat{b}$ is well controlled when sample size is large enough, especially when $r = 5$.

3.6.3 Discussion

The method using order statistics provides plausible estimates provided the sample size is large enough. The estimation is especially outstanding for exponential mixture distributions with medium separation between the components (for instance, when $r = 5$). The estimation of b is poor for samples of small sizes; the resulting variance of the estimator $\hat{b}$ is extremely large, especially when the component distributions are well separated. It is likely to provide unreasonable parameter estimates, either with negative values or in complex forms.

3.7 Comparison of Estimation Methods

We have discussed six different methods, the MLE via the EM algorithm, the method of ordinary moments (MM), the method of fractional moments (FM), the method of attenuated moments (AM), the method of Appell-Fourier Moments with $\alpha = 3$, 4 and 5 (AP_3, AP_4, AP_5), and the method using order statistics (OS), for estimating a mixture of two exponential distributions in this chapter. For each method, we carried out simulation experiments (with 10000 replications) to examine the behaviour of the method for samples with various

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1257	0.1148	0.1110	0.1004	0.1001
$E\left[\hat{b}\right]$	0.1413	0.3432	0.3990	0.5658	0.5130
$E[\hat{p}]$	0.6809	0.6448	0.6340	0.5838	0.5989
$(\hat{a}-a)^2$	0.0007	0.0002	0.0001	1.86×10^{-7}	2.64×10^{-9}
$\left(\hat{b}-b\right)^2$	0.1287	0.0246	0.0102	0.0043	0.0002
$(\hat{p}-p)^2$	0.0065	0.0020	0.0012	0.0003	1.27×10^{-6}
$Var[\hat{a}]$	0.0082	0.0049	0.0035	0.0013	4.79×10^{-5}
$Var\left[\hat{b}\right]$	806	380	465	78	0.0064
$Var[\hat{p}]$	0.1449	0.1460	0.0972	0.0606	0.0031
$MSE[\hat{a}]$	0.0089	0.0051	0.0036	0.0013	4.79×10^{-5}
$MSE\left[\hat{b}\right]$	807	380	465	78	0.0065
$MSE[\hat{p}]$	0.1515	0.1480	0.0983	0.0608	0.00310

Table 3.52: Performance of the method using order statistics for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1248	0.1110	0.1071	0.1023	0.1002
$E\left[\hat{b}\right]$	0.8859	0.9333	0.8080	1.1591	1.0277
$E[\hat{p}]$	0.6116	0.5939	0.5899	0.5945	0.6002
$(\hat{a}-a)^2$	0.0006	0.0001	5.07×10^{-5}	5.33×10^{-6}	2.66×10^{-8}
$\left(\hat{b}-b\right)^2$	0.0130	0.0045	0.0369	0.0253	0.0008
$(\hat{p}-p)^2$	0.0001	3.68×10^{-5}	0.0001	3.03×10^{-5}	5.59×10^{-8}
$Var[\hat{a}]$	0.0107	0.0043	0.0028	0.0008	3.39×10^{-5}
$Var\left[\hat{b}\right]$	6910	7167	8985	3039	0.0241
$Var[\hat{p}]$	0.0962	0.0746	0.0586	0.0258	0.0013
$MSE[\hat{a}]$	0.0113	0.0045	0.0029	0.0008	3.39×10^{-5}
$MSE\left[\hat{b}\right]$	6910	7167	8985	3039	0.0249
$MSE[\hat{p}]$	0.0963	0.0747	0.0587	0.0259	0.00129

Table 3.53: Performance of the method using order statistics for 10000 data sets simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1.0$ and $p = 0.6$ for different sample size n_o.

$r = 2$	Bias^2			Variance		
Method	a	b	p	a	b	p
ML	4.79×10^{-8}	0.0014	1.80×10^{-5}	0.0002	0.0304	0.0289
FM	8.80×10^{-9}	0.0592	0.0001	0.0003	14	0.0702
AM	2.42×10^{-6}	9.95×10^{-6}	4.63×10^{-6}	0.0003	2	0.0697
AP_3	6.47×10^{-7}	0.0018	0.0002	0.0003	6	0.0760
AP_4	3.44×10^{-6}	1.60×10^{-6}	0.0014	0.0004	13	0.6143
AP_5	1.21×10^{-5}	0.0038	0.0020	0.0004	18	8
OS	4.72×10^{-5}	0.0005	0.0012	0.0007	4	0.0908

Table 3.54: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 0.6$.

sample size and different separation between the components. In this section, we draw the simulation results from previous sections and make a comparison of these estimation methods based on their measures of error. For exponential mixture distribution, the estimation of all parameters is generally poor when the number of observations in a sample is small; even the MLEs have large variances and are highly biased. The second rate parameter b is especially difficult to estimate; The variance of the estimator $\hat{b}$ is normally large for small samples. To see a better picture of the performance of these methods, we therefore focus on the estimation results from large samples ($n_o = 1000$) for the comparison of these methods. Since the MM is obviously outperformed by the other methods, we exclude it from our comparison here.

Tables 3.54 to 3.56 evaluate the estimation error through the average bias^2 $\left(\left(\overline{\hat{a}} - a\right)^2\right.$, $\left(\overline{\hat{b}} - b\right)^2$ and $\left.\left(\overline{\hat{p}} - p\right)^2\right)$ and the variance ($Var\left[\hat{a}\right]$, $Var\left[\hat{b}\right]$ and $Var\left[\hat{p}\right]$) over all replications for samples with $r = 2$, 5 and 10 respectively.

Let us first compare the methods for samples with a small difference between the two populations $r = 2$ from Table 3.54. The FM has the smallest bias of $\hat{a}$, the AP_4 has the smallest bias of $\hat{b}$, followed by the AM; whereas the AM has the smallest bias of $\hat{p}$. In terms of the variance, the MLE is the most efficient method for exponential mixture distribution with small separation. Of course, we know that the MLE is asymptotically most efficient so "large" sample sizes must favour the MLE. In general, all moment based method have similar values of $Var\left[\hat{a}\right]$ (about twice the variance of MLE), except from the method using order statistics. The AM has the variances of $\hat{b}$ and $\hat{p}$ which are nearest to the variances given by the MLE. Notably, the variance of $\hat{b}$ given by OS is quite close to the one from AM. The variance of $\hat{p}$ provided by the FM is indeed quite close to the one from AM. As mentioned before, the AP_4 and AP_5 are poor methods for p; the $Var\left[\hat{p}\right]$ of these two methods are unacceptably large, especially AP_5.

Next, we focus on Table 3.55 to study the behaviour of these estimators when they are used to estimate exponential mixture distribution with $r = 5$. Again, the AM has the lowest bias of $\hat{a}$ and $\hat{p}$; whereas the AP_3 provides estimates of b with the lowest bias. The MLE

$r = 5$	Bias^2			Variance		
Method	a	b	p	a	b	p
ML	3.78×10^{-8}	7.36×10^{-5}	1.05×10^{-8}	3.79×10^{-5}	0.0043	0.0021
FM	2.41×10^{-8}	0.0002	9.55×10^{-8}	5.27×10^{-5}	0.0079	0.0037
AM	8.52×10^{-11}	4.00×10^{-5}	2.54×10^{-9}	4.57×10^{-5}	0.0058	0.0030
AP_3	2.18×10^{-10}	7.47×10^{-6}	5.76×10^{-8}	7.01×10^{-5}	0.0164	0.0049
AP_4	4.62×10^{-8}	0.0018	7.38×10^{-5}	7.61×10^{-5}	0.0774	0.1913
AP_5	2.16×10^{-9}	0.0054	0.00242	8.94×10^{-5}	1.0180	3
OS	2.64×10^{-9}	0.0002	1.27×10^{-6}	4.79×10^{-5}	0.0064	0.0031

Table 3.55: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$.

$r = 10$	Bias^2			Variance		
Method	a	b	p	a	b	p
ML	1.44×10^{-8}	3.37×10^{-5}	4.02×10^{-8}	2.46×10^{-5}	0.0089	0.0007
FM	1.60×10^{-9}	0.0001	5.89×10^{-8}	3.21×10^{-5}	0.0177	0.0012
AM	2.49×10^{-11}	2.41×10^{-5}	1.31×10^{-10}	2.84×10^{-5}	0.0116	0.0009
AP_3	3.01×10^{-10}	0.0012	1.19×10^{-8}	5.05×10^{-5}	0.0429	0.0016
AP_4	3.04×10^{-9}	0.0079	0.0079	5.75×10^{-5}	0.1887	0.0571
AP_5	2.84×10^{-8}	0.0044	0.0046	6.67×10^{-5}	11	3
OS	2.66×10^{-8}	0.0008	5.59×10^{-8}	3.39×10^{-5}	0.0241	0.00129

Table 3.56: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$.

is the most efficient method, in terms of the variance, followed by the AM estimator. The OS estimator can be considered as performing equally well as the AM because its variances are only slightly larger than the ones of the AM. Notably, the estimates provided by the method based on Appell sequences are not plausible. The variances of $\hat{b}$ and $\hat{p}$ given by the AP_5 is the largest among all the estimators.

When the separation between the populations are large ($r = 10$), as seen in Table 3.56, the AM estimator is the best method in terms of the bias^2. All methods return plausible estimates of a when both the separation and the sample size are large enough. In term of the variance, the MLE remains as the most efficient method. The AM estimator performs equally well although its variances are larger than the ones given by the MLE in a small margin. The FM estimator and the OS estimator have similar variances in this case. The AP_3 provides reasonable estimates, although its variance of the estimator $\hat{b}$ is considerably large compared to the AM. The AP_4 and AP_5 are under-performed when they are used to estimate $\hat{b}$ and $\hat{p}$. They have relatively higher bias and variances compared their rivals.

Figure 3.24 shows the distribution of various estimators $\hat{a}$ for 1000 observations arising from a mixture of two exponential distributions with $r = 5$ over all replications. All estimators have means close to the true value 0.1. The MLE has the lowest variance, followed

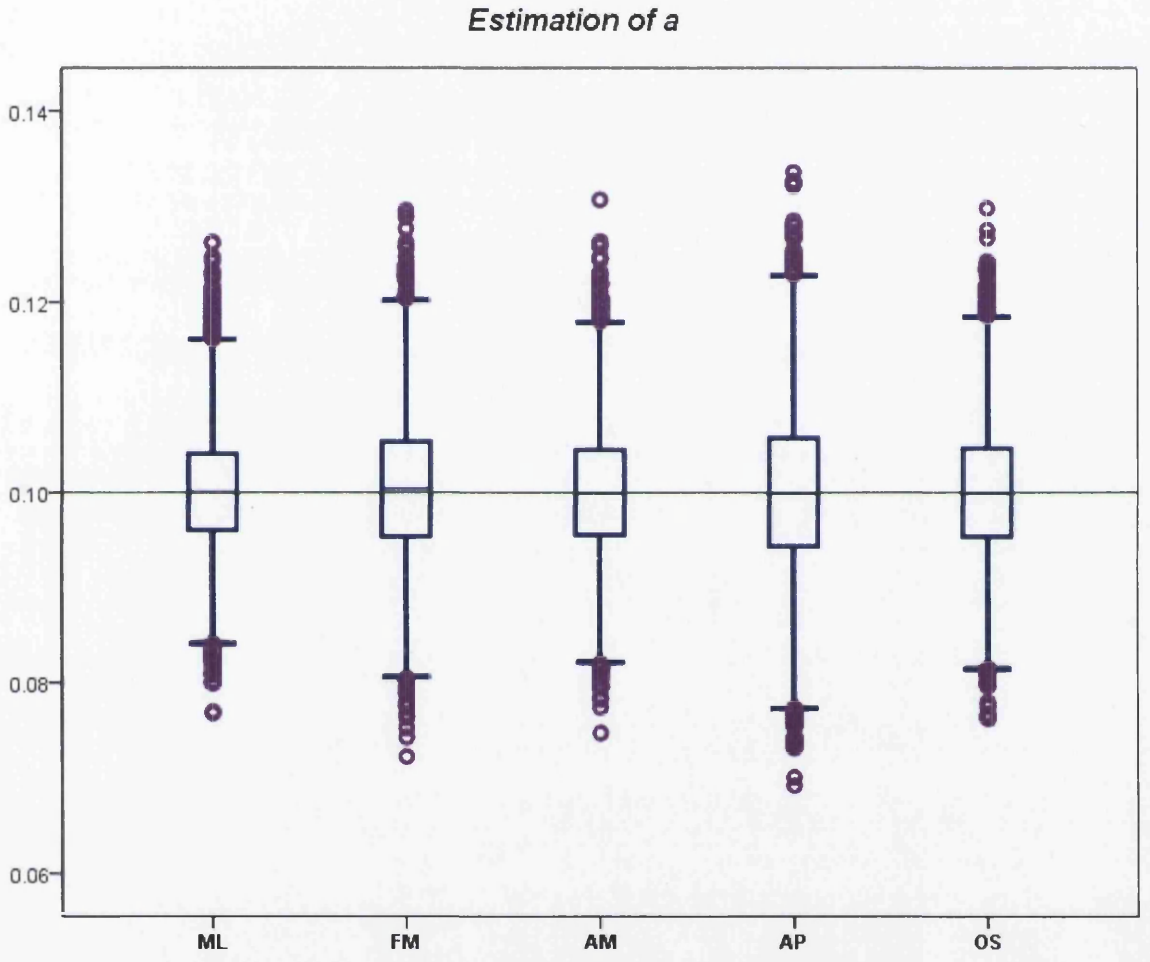


Figure 3.24: Distribution of various estimators $\hat{a}$ for 1000 observations arising from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$. Simulated figures are based on 10000 replications.

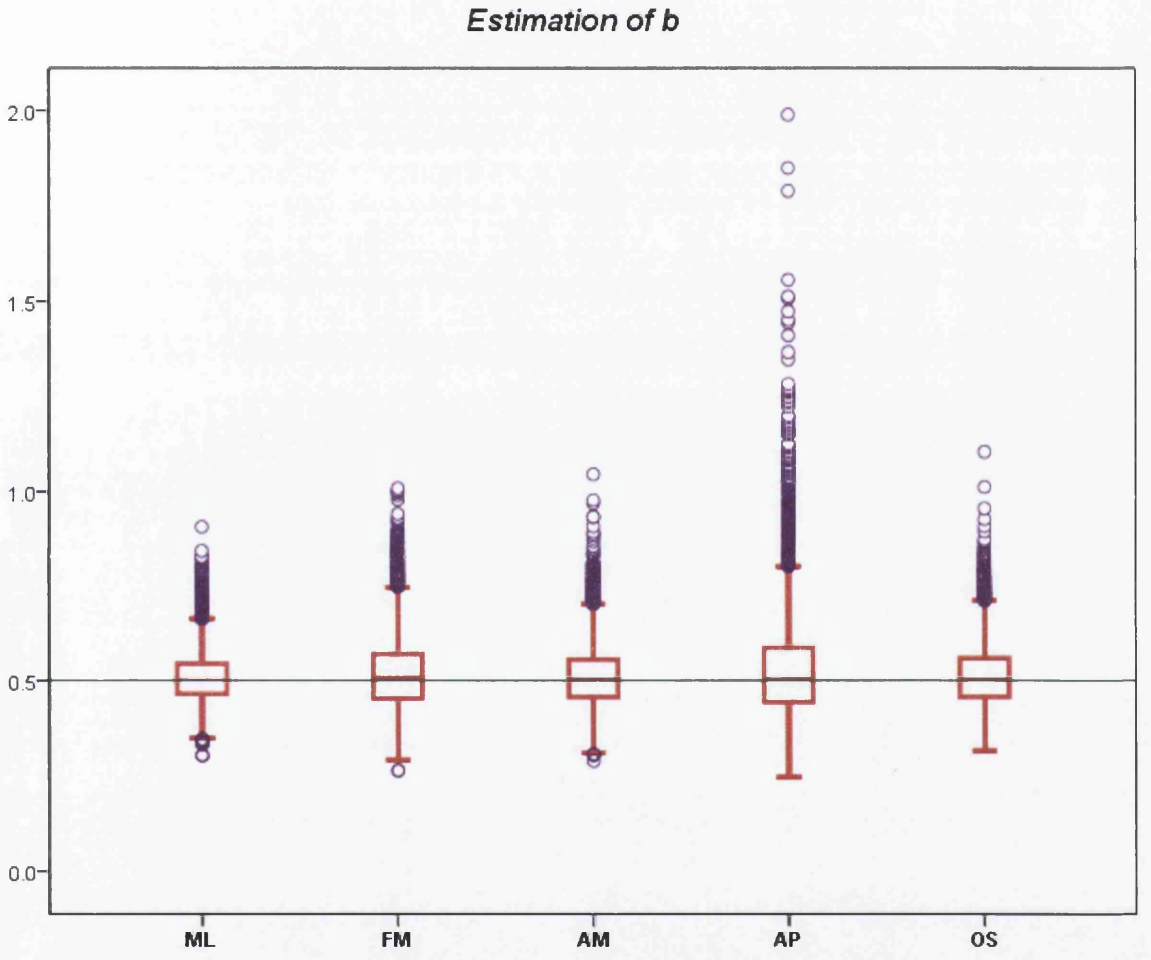


Figure 3.25: Distribution of various estimators $\hat{b}$ for 1000 observations arising from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$. Simulated figures are based on 10000 replications.

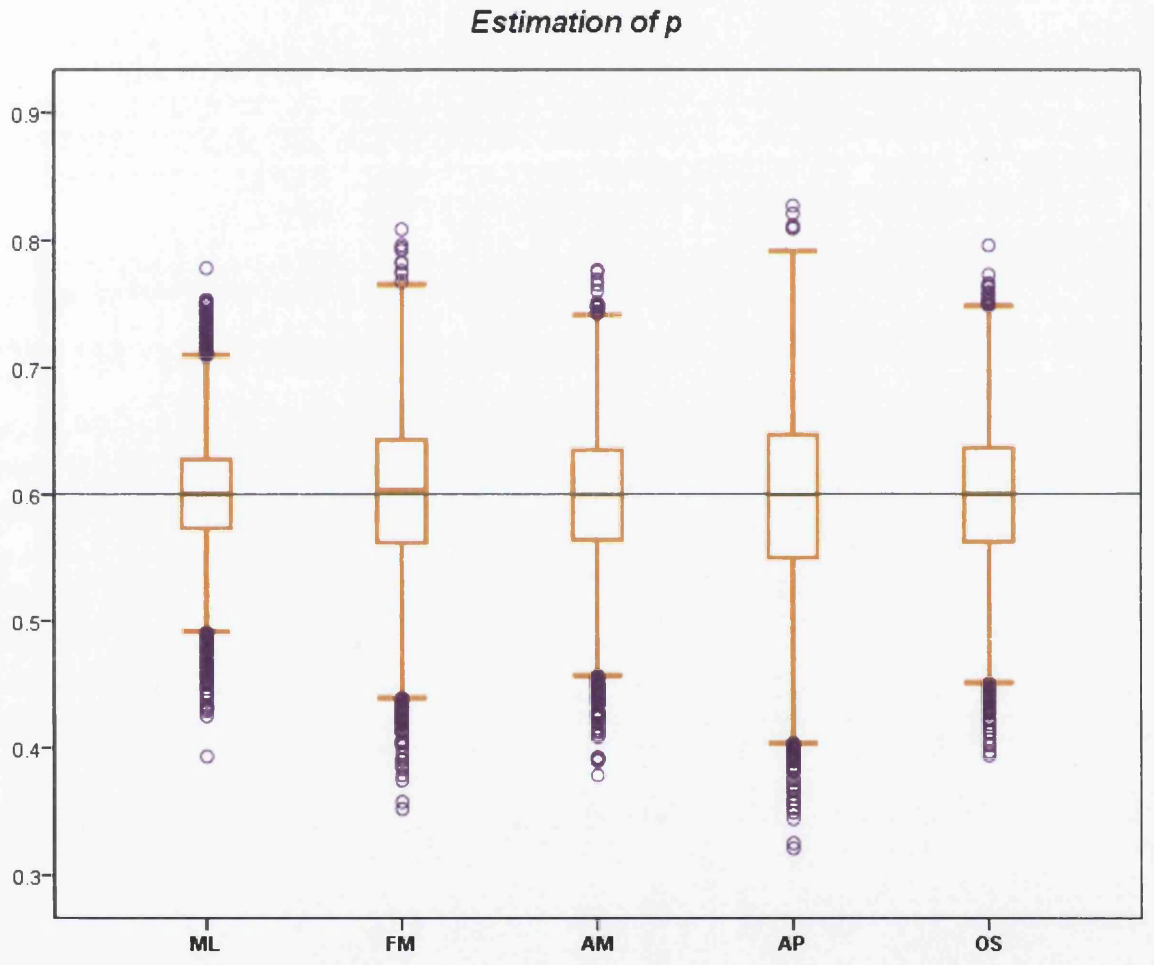


Figure 3.26: Distribution of various estimators $\hat{p}$ for 1000 observations arising from a mixture of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 0.6$. Simulated figures are based on 10000 replications.

$Eff\left[\hat{\Theta}\right]$	$r=2$			$r=5$			$r=10$		
Method	a	b	p	a	b	p	a	b	p
ML	1.8822	0.00144	4.0289	1.1580	1.1330	1.3072	1.0910	1.1273	1.1528
FM	0.9714	3.13×10^{-6}	1.6577	0.8328	0.6179	0.7500	0.8361	0.5684	0.7094
AM	0.9126	2.19×10^{-5}	1.6703	0.9603	0.8458	0.9305	0.9450	0.8649	0.8925
AP_3	0.8857	7.31×10^{-6}	1.5324	0.6261	0.2986	0.5625	0.5315	0.2339	0.5321
AP_4	0.8139	3.38×10^{-6}	0.1896	0.5767	0.0634	0.0144	0.4668	0.0532	0.0145
AP_5	0.8139	2.44×10^{-6}	0.0146	0.4909	0.0049	0.0009	0.4024	0.0009	0.0003
OS	0.4495	1.10×10^{-5}	1.2830	0.9162	0.7726	0.8855	0.7917	0.4165	0.6429

Table 3.57: Efficiencies of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a mixture of two exponential distributions with $a = 0.1$, $b = 0.1r$ and $p = 0.6$.

by the AM and then the OS. Figure 3.25 presents the distribution of different estimators $\hat{b}$ for samples of the same size and distribution. Once again, the variance of the MLE $\hat{b}$ is the smallest; whereas the variances of the AM estimator and the OS estimator are both small and near to the variance of the MLE. The variance of the FM estimator $\hat{b}$ is about twice the one of the MLE. The AP_3 is outperformed by the other methods in estimating b due to its relatively larger variance of the estimator. The distribution of $\hat{p}$ are shown in Figure 3.26. The variance of the AM estimator $\hat{p}$ is the lowest among all the moment-based estimators, and is close to the best estimator ML. The variance of the OS estimator is indeed very small and close to the AM.

Since the Fisher information matrix for a mixture of two exponential distributions can be obtained with Jalali's (2008) solutions, as explained in Section 3.1.6, we shall now find the asymptotic efficiency of each estimator, denoted as $Eff\left[\hat{\Theta}\right]$, by dividing the CRLB of $Var\left[\hat{\Theta}\right]$, presented in Table 3.5, by the simulated variances of the estimators in Tables 3.54 to 3.56; the closer is $Eff\left[\hat{\Theta}\right]$ to one, the more efficient is an estimator. The efficiencies of $\hat{a}$, $\hat{b}$ and $\hat{p}$ for all three degrees of separation ($r = 2$, 5 and 10) are presented in Table 3.57. Note that some efficiencies in the table are greater than one because the Fisher information obtained is only an approximation. Once again, from this table, we confirm that the AM, in most cases, has the highest efficiencies among all moment-based estimators. The MLE is, as expected, the most efficient method; whereas the AP_4 and AP_5 have the least efficiencies in all cases. The FM has satisfactory efficiencies for $\hat{a}$ and $\hat{p}$, especially when $r = 2$; whereas the OS has outstanding efficiencies for all parameters when $r = 5$. All methods have extremely low efficiencies of $\hat{b}$ when the separation between populations is unclear.

To summarise, for "large" samples the MLE, unsurprisingly, appears to be the best method for the estimation problem of exponential mixture distribution. Excitingly, the new methods, the method of attenuated moments and the method using order statistics, perform well in the estimation by returning highly precise estimates with low variances. The method of fractional moments provides reasonable parameter estimates although its variances of

estimators are larger than the AM and the OS in a small margin. The method based on Appell sequences is outperformed by the other methods; the AP_3 will still provide reasonable estimates for samples of large sizes with large difference between the two populations. The AP_4 and AP_5 are poor in estimating b and p, even when the sample size is large. For medium size samples, one cannot say which method is universally better, so having all methods in our disposal could be very helpful.

3.8 Summary

This chapter has concentrated largely on the problem of estimating the parameters in exponential mixture distribution. We first considered the standard approaches of moments and maximum likelihood in estimating the mixture parameters, followed by the investigation of four moment based methods, namely the fractional moment estimator, the attenuated moment estimator, the Appell moment estimator and a method using order statistics. All these modified moment estimators appear to possess much greater efficiency than the ordinary moment estimator. *In particular, the method of attenuated moments outperforms the other moment methods by providing estimates with both relatively lower bias and lower variance. In fact, this method is comparable to the popular MLE; the bias of estimates given by this method are lower than the ML estimates, whereas the variances of the estimators are marginally larger than the ones given by the MLE. The method of attenuated moments provides a quick but accurate approach for users to estimate the parameters in a mixture of exponential densities. Unlike the MLE in which the likelihood equations must be solved using iterative techniques, one simply needs a calculator or an Excel spreadsheet to carry out the calculation to solve the estimation problem.*

We also devised the asymptotic covariance matrix of estimator for three methods, namely the method of fractional moments, the method of attenuated moments and the method of Appell moments. Given these matrices, we are able to suggest, respectively, the optimal fraction, combination of fraction and attenuation and ω for users when using these methods so that the most precise estimates are obtained. We have also suggested a useful way to identify the best estimates given by the fractional moment estimator and the attenuated moment estimator, when the separation between the two components is unknown (this will be so in many practical cases).

It is worth mentioning that the method using order statistics is outstanding in fitting large samples. To conclude, the estimation methods discussed in this chapter should provide reasonable parameter estimates provided the sample size is large enough.

Chapter 4

Mixtures of Geometric Distribution

In Chapter 1, we studied a model in which one state of a Markov chain is isolated while the rest of the states are clumped into one "level" (clump model), with transition matrix

$$\begin{bmatrix} \alpha & \boldsymbol{u} \\ \boldsymbol{v} & \boldsymbol{P} \end{bmatrix}, \tag{4.1}$$

where α is a scalar probability not equal to 1, $\boldsymbol{u}$ is a row m-vector of probabilities, $\boldsymbol{v}$ is a column m-vector of probabilities, and $\boldsymbol{P}$ is an m by m matrix. The discrete lifetime in the level from the time the level is entered into from state 1 is denoted as N and the PMF is given by

$$f(n) = \frac{\boldsymbol{u}}{1-\alpha}\boldsymbol{P}^{n-1}\boldsymbol{v}, \tag{4.2}$$

where $\dfrac{\boldsymbol{u}}{1-\alpha}$ is the probability vector of entry into the level. The generating function of N is

$$G(s) = \frac{\boldsymbol{u}s}{1-\alpha} \times \frac{\boldsymbol{I}-\boldsymbol{P}}{\boldsymbol{I}-s\boldsymbol{P}} \times \mathbf{1}_m. \tag{4.3}$$

We have shown that, when $m = 2$, the discrete waiting time N has a mixture of two geometric distributions.

In the past, several authors have considered discrete mixture distributions, in particular the binomial mixtures by Blischke (1964) and the Poisson mixtures by Rider (1961) and Hasselblad (1969). However, there appears to be little discussion in the literature on geometric mixtures. As shown in Chapter 1, the mixture of geometric distributions is the discrete analogue of the continuous exponential mixtures. This kind of distribution plays important roles in fitting discrete waiting times. Theorem 2 confirms that the moment based methods developed for estimating parameters of the continuous mixed exponential distribution can be adapted easily to the discrete analogue, the mixed geometric distribution. In this chapter, we demonstrate how the methods discussed in the previous chapter are

used to estimate the parameters of N, where N is a discrete random variable arising from a mixture of geometric distribution. We concentrate on the mixture with two components; however the methods studied here can be extended to mixture distributions with more than two components.

If the probability that the n^{th} trial is the first success can be expressed as

$$f(n;\boldsymbol{\Theta}) = pa(1-a)^{n-1} + (1-p)\,b(1-b)^{n-1}, \tag{4.4}$$

where $n = 1,2,...$, $\boldsymbol{\Theta} = (a,b,p)$, $0 \leq a < b \leq 1$ and $0 \leq p \leq 1$, then we say that the distribution of a random variable N is a mixture of two geometric distributions. (4.4) is the PMF of such a distribution; whereas the CDF is defined as

$$F(n;\boldsymbol{\Theta}) = 1 - [p(1-a)^n + (1-p)(1-b)^n]. \tag{4.5}$$

The properties of such a distribution are outlined below:

$$\begin{aligned} E[N] &= \frac{p}{a} + \frac{1-p}{b}, \\ E[N^2] &= p\left(\frac{2-a}{a^2}\right) + (1-p)\left(\frac{2-b}{b^2}\right), \\ Var[N] &= p\left(\frac{2-a}{a^2}\right) + (1-p)\left(\frac{2-b}{b^2}\right) - \left(\frac{p}{a} + \frac{1-p}{b}\right)^2. \end{aligned}$$

In Chapter 3, we let r be the ratio of b to a for a mixture of two exponential distributions with parameter vector $\boldsymbol{\Theta} = (a,b,p)$. In other words, the rate parameter of the second exponential densities can be expressed as $b = ra$. However for a mixture of two geometric distributions, the relationship between a and b is not as straightforward as its continuous analogue. For a mixture of two geometric distributions, we use, for the purpose of investigating mixtures with different degrees of separation between the components,

$$b = 1 - \bar{a}^r, \tag{4.6}$$

where $\bar{a} = 1 - a$. This is equivalent to

$$r = \frac{\ln \bar{b}}{\ln \bar{a}} = \frac{b - \frac{b^2}{2} + \frac{b^3}{3}...}{a - \frac{a^2}{2} + \frac{a^3}{3}...} \approx \frac{b}{a},$$

where $\bar{b} = 1 - b$. For example if we want the ratio r to be 2 when $a = 0.1$, then

$$b = 1 - (1 - 0.9)^2 = 0.19.$$

In Figure 4.1 we see the PMF plot of a mixture of geometric distributions where the component distributions are not well separated ($r = 2$), alongside the PMF plots of its two components. The mixture PMF looks similar to its components and this makes the parame-

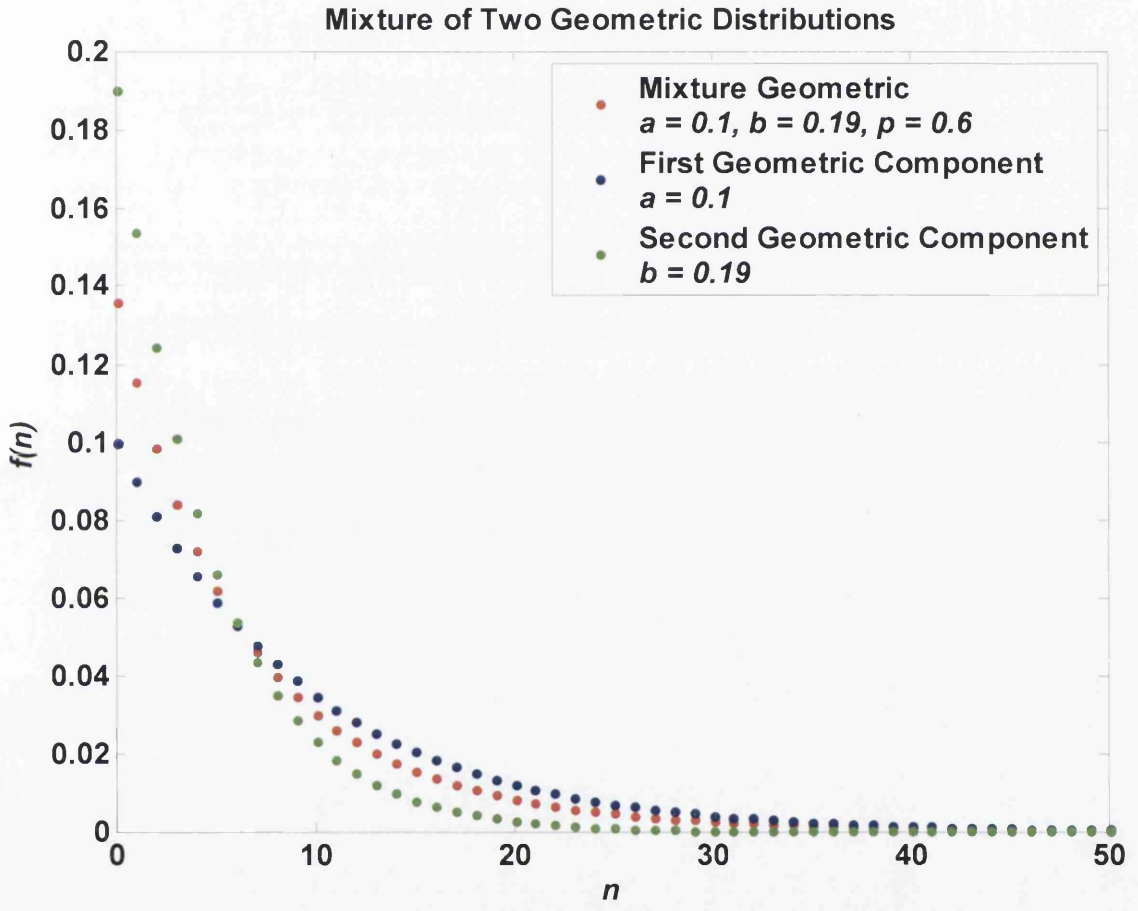


Figure 4.1: PMF plot of a mixture of two geometric distributions together with the PMF plots of its components: $a = 0.1$, $b = 0.19$ and $p = 0.6$.

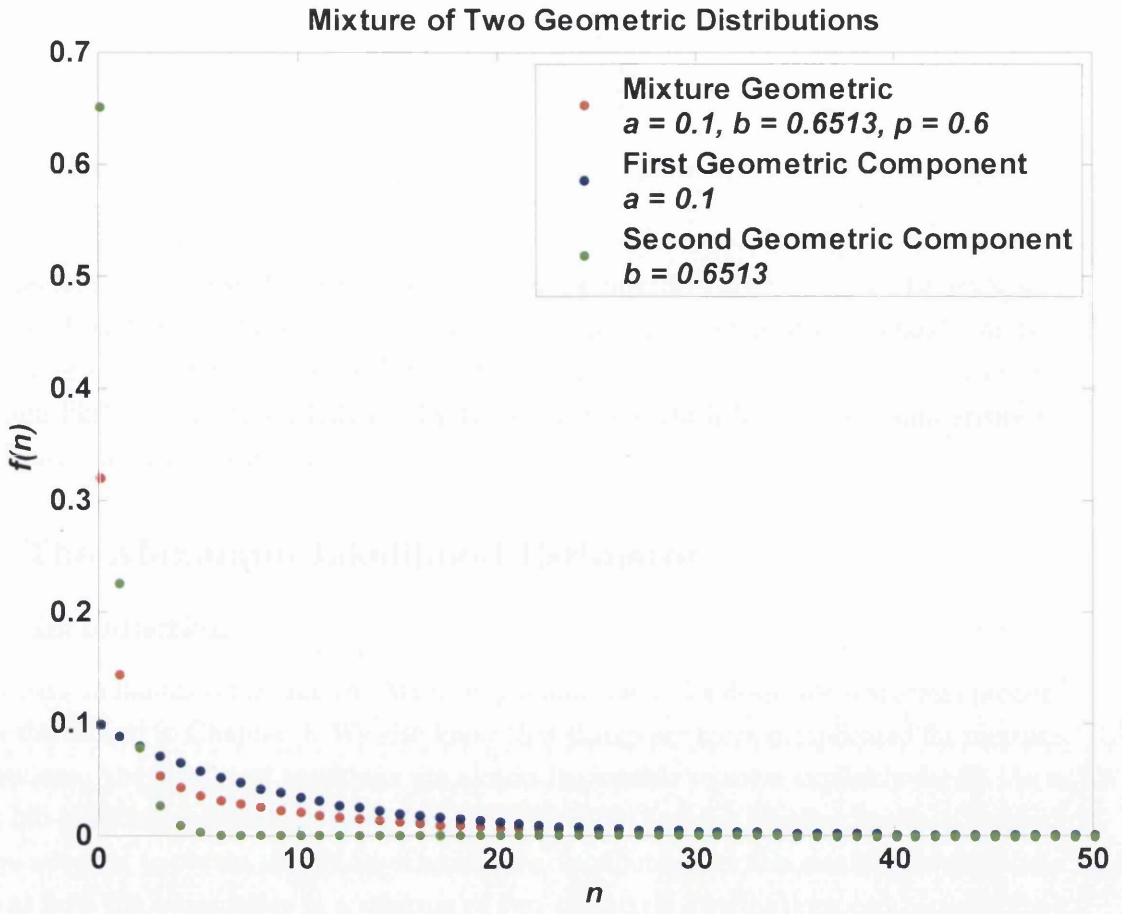


Figure 4.2: PMF plot of a mixture of two geometric distributions together with the PMF plots of its components: $a = 0.1$, $b = 0.6513$ and $p = 0.6$.

ter estimation of such a mixture difficult, essentially because it is hardly distinguishable from a single distribution. Figure 4.2 shows the PMF plot of a well separated two-component geometric mixture with $r = 10$ and the PMF plots of its components. Compared to the first component, the PMF plot of the second component is more skewed, indicating that shorter discrete waiting time is significantly more probable than longer waiting time; whereas the mixture PMF plot shows a blend of the characteristics of both components.

The degree of separation between component distributions has a significant effect on the performance of any estimation technique in the analysis of mixed samples. Throughout this thesis, we consider mixtures of geometric distributions with different ratios r, where a is fixed at 0.1, representing separation ranging from small ($r = 2$) over medium ($r = 5$) to large ($r = 10$). Table 4.1 summarises the true values of the parameters in these cases; whereas Figure 4.3 shows the PMF plots of these mixtures. It is clear from the figure that the PMF plot is more skewed for a mixture with a larger separation between the

r	a	b	p
2	0.1	0.1900	0.6
5	0.1	0.4905	0.6
10	0.1	0.6513	0.6

Table 4.1: True values of parameters of mixtures of geometric distributions with respect to different ratio r.

components. It is also worth noting that all geometric mixtures are unimodal like any single geometric distribution. We shall now move on to study a few estimation methods for this discrete mixture distribution in the following sections, beginning with the most well known maximum likelihood method, followed by three methods which have formal similarities to the ordinary moment estimator.

4.1 The Maximum Likelihood Estimator

4.1.1 Introduction

The maximum likelihood estimator (MLE) is popular due to its desirable statistical properties, as mentioned in Chapter 3. We also know that things are more complicated for mixture distributions: the likelihood equations are almost impossible to solve explicitly for $\boldsymbol{\Theta}$. As a result, hill-climbing techniques such as the EM algorithm and the Newton Raphson method must be adopted to obtain the MLEs of a mixture distribution. In this section, we shall take a look at how the parameters in a mixture of two geometric distributions can be estimated using the MLE via the EM algorithm.

4.1.2 The Expectation-Maximisation Algorithm

In Section 3.1.3 we have demonstrated the application of the EM algorithm for the ML fitting of a mixture of two exponential distributions. In this section, we apply this algorithm to the discrete analogue, a mixture of two geometric distributions, with a similar procedure. With the addition of the hidden component label data z_{ji} to the problem, the likelihood function is considered as complete when it is expressed in the form of

$$l_c(\boldsymbol{\Theta}) = \sum_{i=1}^{n_o} z_{1i} \left[\log p + \log a + (n_i - 1)\log(1-a)\right] + z_{2i} \left[\log(1-p) + \log b + (n_i - 1)\log(1-b)\right].$$

The conditional expected complete data log-likelihood $Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)}\right)$ is then computed in the E-step; whereas in the M-step, the estimates of parameter $\boldsymbol{\Theta}$ is updated by maximising $Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)}\right)$.

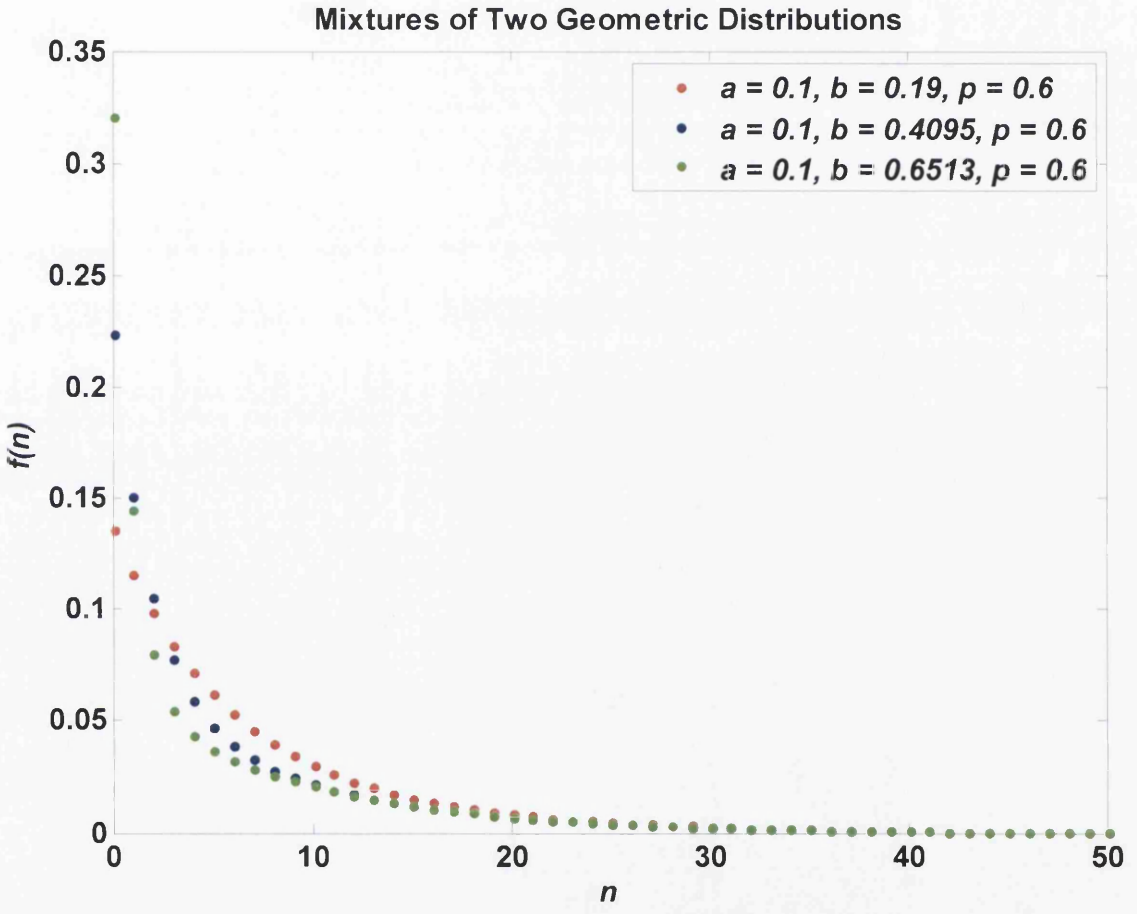


Figure 4.3: PMF plots of mixtures of two geometric distributions for varying separation.

E-step

The posterior probabilities are calculated using (3.14):

$$\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)=\frac{p^{(k)}a^{(k)}\left(1-a^{(k)}\right)^{n_i-1}}{p^{(k)}a^{(k)}\left(1-a^{(k)}\right)^{n_i-1}+\left(1-p^{(k)}\right)b^{(k)}\left(1-b^{(k)}\right)^{n_i-1}}, \tag{4.7}$$

$$\tau_2\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)=\frac{\left(1-p^{(k)}\right)b^{(k)}\left(1-b^{(k)}\right)^{n_i-1}}{p^{(k)}a^{(k)}\left(1-a^{(k)}\right)^{n_i-1}+\left(1-p^{(k)}\right)b^{(k)}\left(1-b^{(k)}\right)^{n_i-1}}. \tag{4.8}$$

Following (3.15), the conditional expected complete data log-likelihood is in the form of

$$Q\left(\boldsymbol{\Theta};\hat{\boldsymbol{\Theta}}^{(k)}\right)=\sum_{i=1}^{n_o}\left\{\begin{array}{c}\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)\left[\log p+\log a+(n_i-1)\log(1-a)\right]\\+\tau_2\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)\left[\log(1-p)+\log b+(n_i-1)\log(1-b)\right]\end{array}\right\}. \tag{4.9}$$

where $\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)$ and $\tau_2\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)$ are given by (4.7) and (4.8).

M-step

The mixing probability $\hat{p}$ is estimated using the posterior probabilities in (4.7)

$$\hat{p}^{(k+1)}=\sum_{i=1}^{n_o}\frac{\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)}{n_o}. \tag{4.10}$$

The M-step for geometric components exists in closed form. Recalling (3.22), $\hat{a}^{(k+1)}$ is simply the root of

$$\sum_{i=1}^{n_o}\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)\frac{\partial}{\partial a}\left[\log p+\log a+(n_i-1)\log(1-a)\right]=0$$

$$\Leftrightarrow\sum_{i=1}^{n_o}\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)\left[\frac{1}{a}-\frac{n_i-1}{1-a}\right]=0.$$

Therefore, $\boldsymbol{\theta}=(a,b)$ is updated by

$$\hat{a}^{(k+1)}=\frac{\sum_{i=1}^{n_o}\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)}{\sum_{i=1}^{n_o}n_i\tau_1\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)}. \tag{4.11}$$

Similarly,

$$\hat{b}^{(k+1)}=\frac{\sum_{i=1}^{n_o}\tau_2\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)}{\sum_{i=1}^{n_o}n_i\tau_2\left(n_i;\hat{\boldsymbol{\Theta}}^{(k)}\right)}. \tag{4.12}$$

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1193	0.1137	0.1095	0.1033	0.0998
$E[\hat{b}]$	0.2561	0.2484	0.2515	0.2415	0.2145
$E[\hat{p}]$	0.5853	0.5843	0.5803	0.5728	0.6049
$(\hat{a}-a)^2$	0.0004	0.0002	9.03×10^{-5}	1.07×10^{-5}	5.97×10^{-8}
$(\hat{b}-b)^2$	0.0044	0.0034	0.0038	0.0027	0.00060
$(\hat{p}-p)^2$	0.0002	0.0003	0.0004	0.0007	2.40×10^{-5}
$Var[\hat{a}]$	0.0026	0.0018	0.0014	0.0009	0.0002
$Var[\hat{b}]$	0.0489	0.0440	0.0436	0.0365	0.0076
$Var[\hat{p}]$	0.0182	0.0222	0.0260	0.0378	0.0282
$MSE[\hat{a}]$	0.0030	0.0019	0.0015	0.0009	0.0002
$MSE[\hat{b}]$	0.0533	0.0474	0.0474	0.0391	0.0082
$MSE[\hat{p}]$	0.0185	0.0225	0.0264	0.0385	0.0282

Table 4.2: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 0.6$ for different sample size n_o. Starting values are set as true values.

where $\tau_1\left(n_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)$ and $\tau_2\left(n_i; \hat{\boldsymbol{\Theta}}^{(k)}\right)$ are updated according to (4.7) and (4.8) respectively. The iteration is repeated until the stopping criterion suggested by Böhning *et al.* (1994) using Aitken's acceleration (as described in Section 3.1.3') is satisfied.

4.1.3 Simulation Results

For illustration, a simulation experiment is carried out in order to examine the performance of the MLE for different sample sizes $n_o = (10, 15, 20, 50, 1000)$ and different separations between the two components. 10000 data sets each consisting of n_o observations were simulated from a two-component geometric mixture model with $a = 0.1$, $b = 1 - 0.9^r$ and $p = 0.6$. The "unknown" parameter vector $\boldsymbol{\Theta} = (a, b, p)$ is estimated for each of the 10000 data sets based on the MLE using the EM algorithm. For simplicity, we use the true values of the parameters as the starting values, $\boldsymbol{\Theta}^{(0)} = (0.1, 1 - 0.9^r, 0.6)$; the stopping criterion adopted is the Aitken's method discussed in (3.32), whereas the tolerance value is set as 0.00001.

Tables 4.2, 4.3 and 4.4 evaluate the estimation error through the average square of bias in estimator $\left(E\left[\hat{\boldsymbol{\Theta}}\right] - \boldsymbol{\Theta}\right)^2$, the variance $Var\left[\hat{\boldsymbol{\Theta}}\right]$ and the mean square error $MSE\left[\hat{\boldsymbol{\Theta}}\right]$ over all replications for a geometric mixture model with $r = 2$, $r = 5$ and $r = 10$ respectively. For each r, $Var[\hat{a}]$ and $Var\left[\hat{b}\right]$ decrease gradually when sample size increases. Nevertheless, when r is small ($r = 2$ and 5), $(\hat{p} - p)^2$ and $Var[\hat{p}]$ actually increase when the sample size increases from $n_o = 10$ to 50. Comparing these three tables, it is clear that, as might be expected, the MLE via the EM algorithm performs best for a mixture of two geometric

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1259	0.1140	0.1103	0.1013	0.1002
$E\left[\hat{b}\right]$	0.4689	0.4726	0.4733	0.4660	0.4138
$E[\hat{p}]$	0.5979	0.5900	0.5855	0.5798	0.5998
$\left(\bar{\hat{a}} - a\right)^2$	0.0007	0.0002	0.0001	1.68×10^{-6}	2.53×10^{-8}
$\left(\bar{\hat{b}} - b\right)^2$	0.0035	0.0040	0.0041	0.0032	1.83×10^{-5}
$\left(\bar{\hat{p}} - p\right)^2$	4.62×10^{-6}	0.0001	0.0002	0.0004	4.51×10^{-8}
$Var[\hat{a}]$	0.0051	0.0029	0.0021	0.0008	3.51×10^{-5}
$Var\left[\hat{b}\right]$	0.0798	0.0727	0.0674	0.0434	0.0017
$Var[\hat{p}]$	0.0392	0.0417	0.0423	0.0355	0.0022
$MSE[\hat{a}]$	0.0058	0.0031	0.0022	0.0008	3.51×10^{-5}
$MSE\left[\hat{b}\right]$	0.0833	0.0766	0.0715	0.0466	0.0017
$MSE[\hat{p}]$	0.0392	0.0418	0.0426	0.0359	0.0022

Table 4.3: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$ for different sample size n_o. Starting values are set as true values.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1222	0.1103	0.1070	0.1023	0.1002
$E\left[\hat{b}\right]$	0.6749	0.6818	0.6862	0.6773	0.6527
$E[\hat{p}]$	0.5905	0.5845	0.5846	0.5883	0.5998
$\left(\bar{\hat{a}} - a\right)^2$	0.0005	0.0001	4.95×10^{-5}	5.06×10^{-6}	3.06×10^{-8}
$\left(\bar{\hat{b}} - b\right)^2$	0.0006	0.0009	0.0012	0.0007	1.95×10^{-6}
$\left(\bar{\hat{p}} - p\right)^2$	9.08×10^{-5}	0.0002	0.0002	0.0001	2.90×10^{-8}
$Var[\hat{a}]$	0.0058	0.0028	0.0017	0.0006	2.32×10^{-5}
$Var\left[\hat{b}\right]$	0.0705	0.0601	0.0513	0.0304	0.00142
$Var[\hat{p}]$	0.0432	0.0382	0.0323	0.0175	0.0008
$MSE[\hat{a}]$	0.0062	0.0029	0.0018	0.0006	2.33×10^{-5}
$MSE\left[\hat{b}\right]$	0.0711	0.0610	0.0525	0.0311	0.0014
$MSE[\hat{p}]$	0.0433	0.0385	0.0325	0.0176	0.0008

Table 4.4: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 0.6$ for different sample size n_o. Starting values are set as true values.

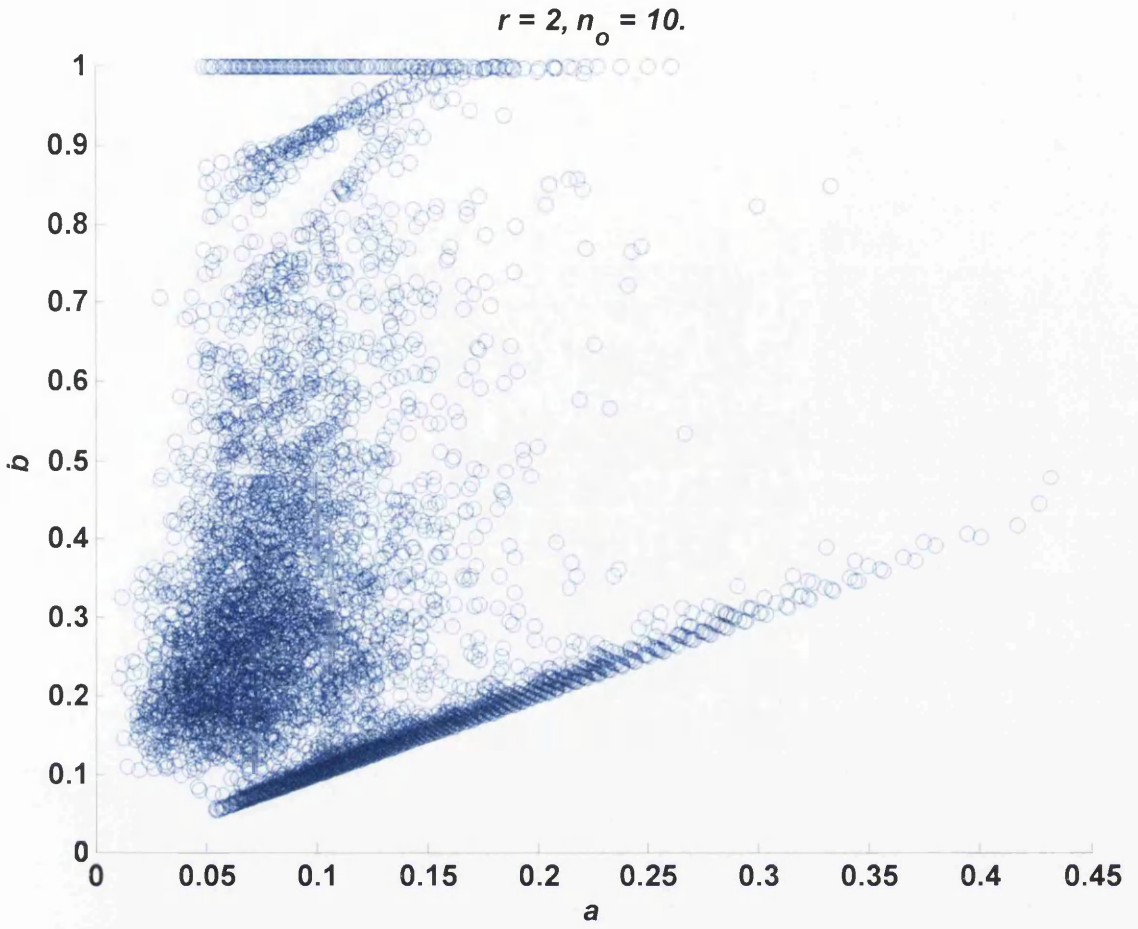


Figure 4.4: Scatter plot of MLE $\hat{b}$ versus $\hat{a}$ for mixtures of two geometric distributions with $a = 0.1$, $b = 0.19$, $p = 0.6$ and $n_o = 10$.

distributions with a large sample size and a large separation between the two components.

Unlike its continuous analogue, mixture of exponentials, the ML estimates have low bias and small variances when n_o is small, even when components are less clearly separated. We recall the performance of the MLE on mixture exponential distribution from Tables 3.1 to 3.3, the variance of the second rate parameter estimate can be very large for small samples. Conversely, $Var\left[\hat{b}\right]$ are small here, even for small samples with a small separation between the components.

We take a closer look on the 10000 estimates from data sets which are small in both size and separation between the components ($n_o = 10$, $r = 2$) by plotting a scatter plot of $\hat{b}$ against $\hat{a}$ in Figure 4.4. We discover that a majority of $\hat{b}$ has a close value to $\hat{a}$, as shown by the straight line at the lower part of the scatter plot. Indeed, 56.63% of the estimates have $\left|\hat{b} - \hat{a}\right| \leq 0.01$, in other words most of the fitted distributions are single geometric distributions rather than mixtures; whereas 2.7% of $\hat{b}$ has a value of 1, as shown

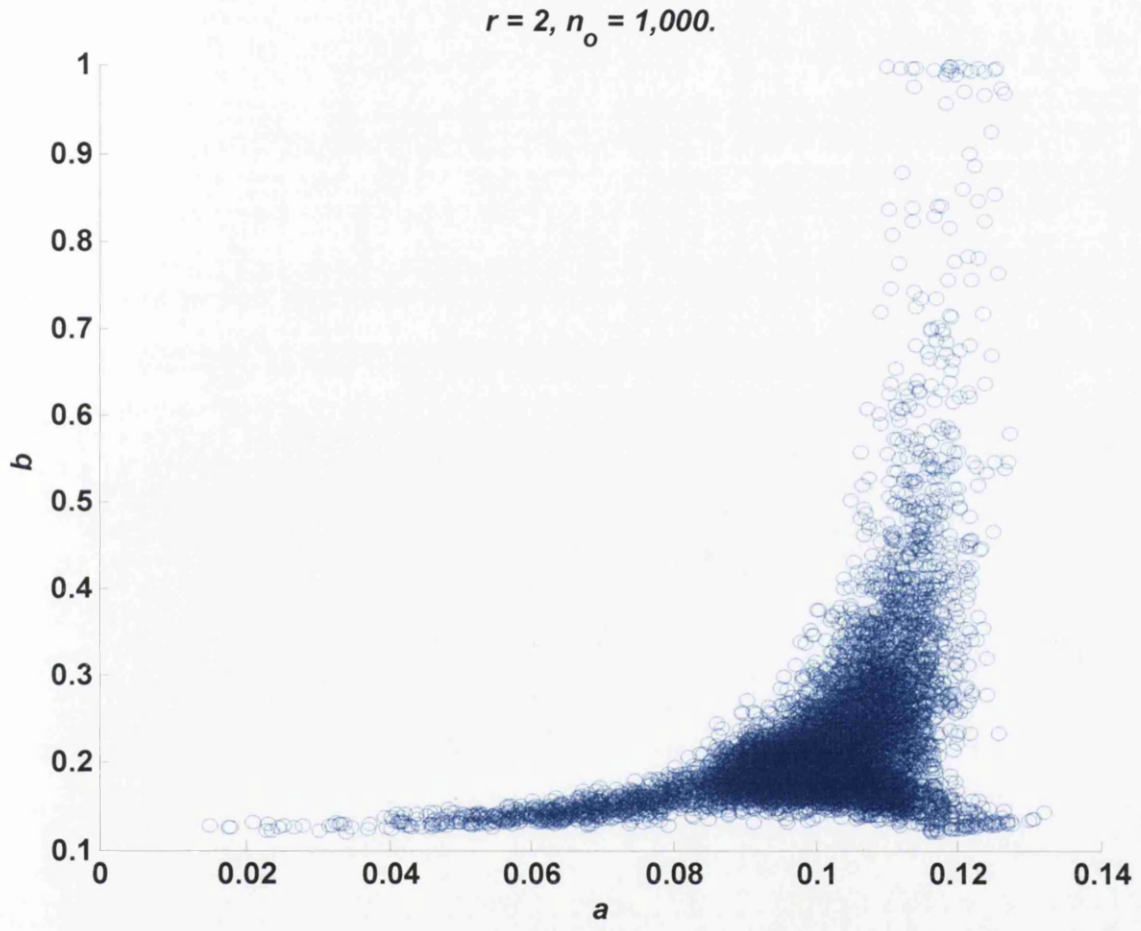


Figure 4.5: Scatter plot of MLE $\hat{b}$ versus $\hat{a}$ for mixtures of two geometric distributions with $a = 0.1$, $b = 0.19$, $p = 0.6$ and $n_o = 1000$.

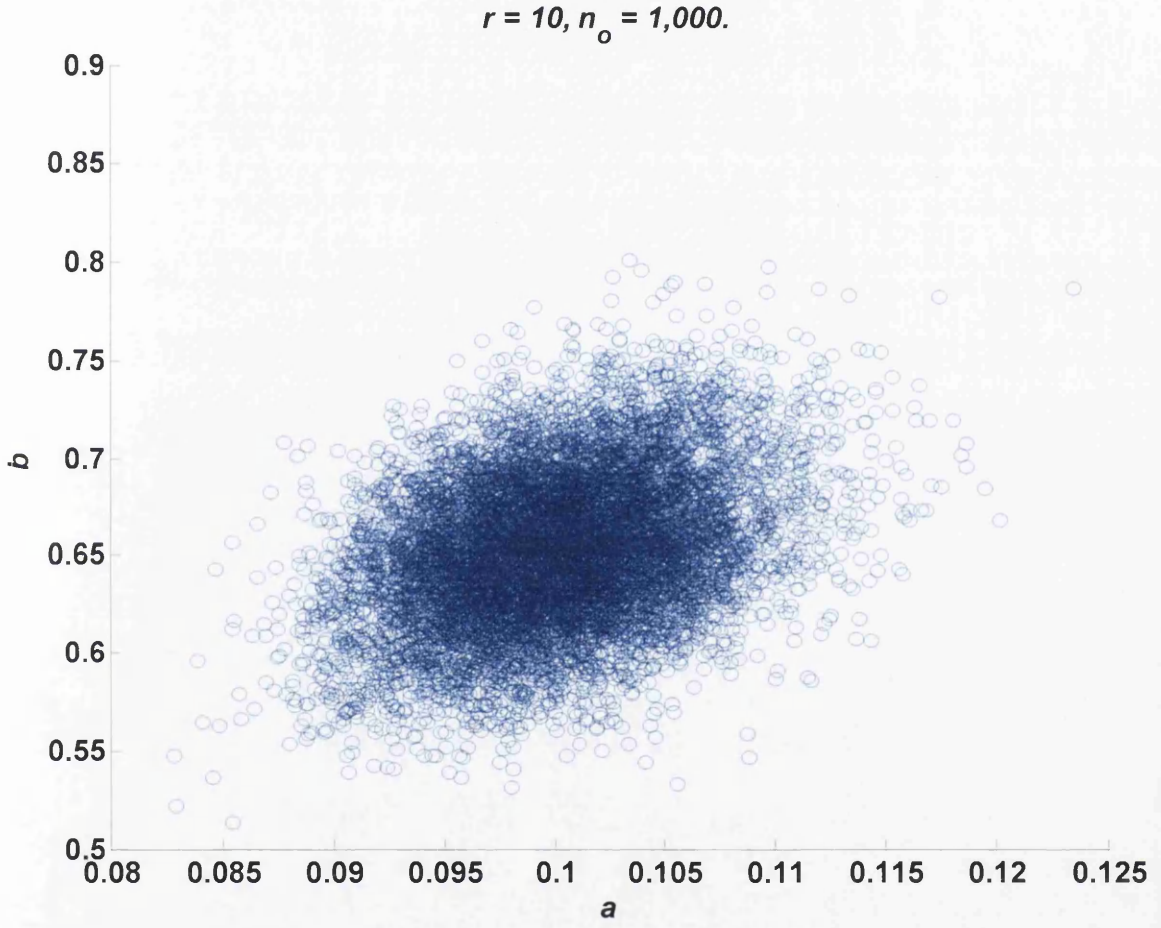


Figure 4.6: Scatter plot of MLE $\hat{b}$ versus $\hat{a}$ for mixtures of two geometric distributions with $a = 0.1$, $b = 0.6513$, $p = 0.6$ and $n_o = 1000$.

by the horizontal line on the upper part of the scatter plot. We chose the initial values to fall within the admissible limits for the parameters (a, b and p are all probabilities with a value between 0 and 1), therefore it is impossible to obtain estimates outside these limits. Since all estimates lie between 0 and 1, their variances are small although most of the fitted distributions have poor goodness of fit for small samples. We show the scatter plot of $\hat{b}$ versus $\hat{a}$ estimated from data sets with the same separation ($r = 2$) but a large number of observations ($n_o = 1000$) in Figure 4.5. The estimation of a and b are clearly improved when the number of observations in a data set is increased, as most of the estimates lie near to the region of the true value ($a = 0.1$ and $b = 0.19$). For mixtures with more clearly separated component densities, the estimation problem is obviously easier especially when the sample size is as large as 1000. We see a scatter plot of $\hat{b}$ versus $\hat{a}$ estimated from data sets, each consisting 1000 observations, arising from a mixture of two geometric distributions with $r = 10$ in Figure 4.6; a great majority of the estimates lie near to the region of the true value ($a = 0.1$ and $b = 0.6513$).

4.1.4 Discussion

For a mixture of two geometric distributions, the ML method should be able to provide reasonable parameter estimates. The variances of the estimators are considerably low, even for samples of small sizes. However, our investigation shows that most of the fitted distributions for small samples are actually a single geometric distribution. Since we know all parameters are probabilities and should lie in the interval $(0, 1)$, we choose the initial values within this limit and hence the final values will always lie inside the admissible limit. As a consequence, the variances of the estimators are small even for small samples, although the ML inferred distributions are not ideal.

As we mentioned in Chapter 3, good initial values are vital when ML method is used. As seen from our simulation results, problems have arisen even when starting values are set as the true values for small samples. With less clearly separated component distributions, various starting values may lead to widely different final estimates. To conclude, for estimating the parameters in a mixture geometric where the two components are not well separated, very large samples may be needed.

4.2 The Method of Rising Factorial Moments

4.2.1 Introduction

Several authors, for example, Blischke (1964), Rider (1961) and Hasselblad (1969) have made use of factorial moments to estimate the parameters in discrete mixtures in which the components are binomial or Poisson. In this section, we apply the method of rising factorial moments on the estimation problem for the distribution of the discrete lifetime, denoted by N, in the level of a hidden Markov chain with transition matrix (4.1). As shown earlier,

when the number of hidden states in the level is $m = 2$, the PMF of N is in the form of (4.4), which is a linear combination (not necessarily a positive mixture) of two geometric distributions. In this chapter, we concentrate on the positive mixture in which the mixing weight p is in the interval $(0, 1)$ before we extend our study to the case when p is allowed to be greater than 1 in the next chapter.

We denote the k^{th} rising factorial moment as ρ_k, defined as

$$\rho_k = E\left[N\left(1+N\right)...\left(\kappa+N-1\right)\right] = E\left[\frac{\Gamma\left(N+k\right)}{\Gamma\left(N\right)}\right]. \tag{4.13}$$

For a two-component geometric mixture model, we consider the first three rising factorial moments

$$\rho_k = k!\left[\frac{p}{a^k} + \frac{(1-p)}{b^k}\right] \tag{4.14}$$

for $k = 1$, 2 and 3. Now if we set

$$z_k = \frac{\rho_k}{k!}, \tag{4.15}$$

i.e. if we choose z_k's to be *normalised* rising factorial moments of our discrete distribution, by analogy with the continuous case, we obtain the same equation for the parameters a, b and p following the procedure from (3.67) to (3.70) in Section 3.2.1. To estimate these parameters, we simply replace z_k's by $\hat{z}_k$'s, where the latter are the estimates of the rising factorial moments from our discrete data, given by

$$\hat{z}_k = \frac{1}{n_o\Gamma\left(k+1\right)}\sum_{i=1}^{n_o}\frac{\Gamma\left(n_i+k\right)}{\Gamma\left(n_i\right)}.$$

So far things are exactly analogous to the continuous case. The parameter $\bar{a} = 1 - a$ is the largest eigenvalue of $\boldsymbol{P}$ in (4.1) which is positive and not exceeding 1. $\bar{b} = 1 - b$ is the other eigenvalue of $\boldsymbol{P}$ which is real and not greater than $\bar{a}$ in absolute value, but may be negative. The fact that this eigenvalue can be negative adds a great deal of subtly to the discrete case. First we note that the PMF in (4.4) can be an actual PMF if it remains everywhere non-negative. This imposes the following restrictions on p:

$$\begin{aligned} p &\leq \frac{1-\bar{b}}{\bar{a}-\bar{b}} \quad \text{if } \bar{b} \geq 0, \\ \frac{-\bar{b}\left(1-\bar{b}\right)}{\left(\bar{a}-\bar{b}\right)} &\leq p \leq \frac{1-\bar{b}}{\bar{a}-\bar{b}} \quad \text{if } \bar{b} < 0. \end{aligned} \tag{4.16}$$

Given any PMF f as in (4.4) with the above constraints on its parameters, we can construct matrices $\boldsymbol{P}$ whose associated N has a PMF equal to f. When $\bar{b} > 0$, the construction will be exactly analogous to the continuous case. Both cases have been thoroughly investigated by Jalali (2002 and 2005c). When $\bar{b} < 0$, we need a different type of construction. The following theorem completes the problem of matrix reconstruction.

r	2	3	4	5	6
$E[\hat{a}]$	0.0999	0.1017	0.1015	0.1015	0.1011
$E[\hat{b}]$	−0.1827	0.3535	0.3524	0.5454	0.7685
$E[\hat{p}]$	0.6375	0.6325	0.6230	0.6200	0.6144
$(\overline{\hat{a}} - a)^2$	1.02×10^{-8}	2.78×10^{-6}	2.15×10^{-6}	2.14×10^{-6}	1.24×10^{-6}
$(\overline{\hat{b}} - b)^2$	0.1389	0.0068	0.0001	0.0185	0.0900
$(\overline{\hat{p}} - p)^2$	0.0014	0.0011	0.0005	0.0004	0.0002
$Var[\hat{a}]$	0.0003	0.0002	0.0001	0.0001	0.0001
$Var[\hat{b}]$	12551	65	621	166	155
$Var[\hat{p}]$	0.0781	0.0306	0.0196	0.0156	0.0138
$MSE[\hat{a}]$	0.0003	0.0002	0.0001	0.0001	0.0001
$MSE[\hat{b}]$	12551	65	621	166	155
$MSE[\hat{p}]$	0.0795	0.0316	0.0201	0.0160	0.0140

Table 4.5: Performance of the method of rising factorial moments for 10000 data sets simulated from a mixture of two geometric distributions with varying $b = 1 - 0.9^r$ and fixed $a = 0.1$ and $p = 0.6$. r ranging from 2 to 6.

4.2.2 Simulation Results

We carried out a simulation experiment to examine the performance of the rising factorial moment estimator in the estimation problem of a two-component geometric mixture model with different level of separation between the two components. As we expect this method to be implausible for samples of small sizes, we focus on large samples by simulating data sets with 1000 observations. In order to investigate the effect of separation between the components on the estimation, we considered nine degrees of separation varying from $r = 2$ to $r = 10$. For each r, we generated 10000 data sets from a mixture of two geometric distributions with $a = 0.1$, $b = 1 - 0.9^r$ and $p = 0.6$, and estimated the parameters using the method of rising factorial moments. For ease of analysis, we exclude any complex estimates during the simulation experiment. The estimation results are presented in Tables 4.5 and 4.6. Judging from $E\left[\hat{\boldsymbol{\Theta}}\right]$ and the bias^2, the estimation of a and p are reasonable as they are close to the true values. Both of the $Var[\hat{a}]$ and $Var[\hat{p}]$, and hence their mean square errors, decrease as the components become further from each other. Unfortunately, the estimation of b, like the continuous analogue, is not consistent; we note the negative average of $\hat{b}$, as shown in the tables, when $r = 2$ and 7. Since $\hat{b}$ is given by $\hat{y}^{-1}$, where $\hat{y}$ is the smaller root of the quadratic equation (3.68), $\hat{b}$ is not necessarily to be found in the interval $(0, 1)$. It may be negative, greater than 1 or in a complex form. We have shown in Chapter 3 that $\hat{b}$ is indeed very sensitive to $\hat{y}$ (see Figure 3.9), a small departure of $\hat{y}$ from its true value can cause an extremely large estimate of b, which is not realistic as b is a probability.

r	7	8	9	10
$E[\hat{a}]$	0.1013	0.1012	0.1011	0.1012
$E[\hat{b}]$	−0.6074	0.8562	0.3460	1.0871
$E[\hat{p}]$	0.6157	0.6144	0.6123	0.6122
$(\bar{\hat{a}} - a)^2$	1.66×10^{-6}	1.40×10^{-6}	1.24×10^{-6}	1.44×10^{-6}
$(\bar{\hat{b}} - b)^2$	1.2749	0.0822	0.07108	0.1899
$(\bar{\hat{p}} - p)^2$	0.0002	0.0002	0.00015	0.0001
$Var[\hat{a}]$	0.0001	0.0001	0.00010	9.76×10^{-5}
$Var[\hat{b}]$	10658	390	824	1232
$Var[\hat{p}]$	0.0119	0.0108	0.0108	0.0100
$MSE[\hat{a}]$	0.0001	0.0001	0.0001	9.90×10^{-5}
$MSE[\hat{b}]$	10659	390	824	1232
$MSE[\hat{p}]$	0.0122	0.0110	0.0109	0.0101

Table 4.6: Performance of the method of rising factorial moments for 10000 data sets simulated from a mixture of two geometric distributions with varying $b = 1 - 0.9^r$ and fixed $a = 0.1$ and $p = 0.6$. r ranging from 7 to 10.

4.2.3 Discussion

The method of moments was the most popular way of estimating the parameters in a mixture distribution before computers became readily available. It is simpler and quicker than the more statistically respectable ML approach. Nevertheless, in these days of very powerful computers, the "highly intractable" likelihood equations can be solved with little difficulty. Many authors (for example, Blischke (1964) and Hasselblad (1969)) had compared the performance of these two approaches and found that the method of moments is outperformed by the MLE in most cases. Hasselblad (1969) showed that the moment estimators for mixtures of binomial distributions have sample variances which are uniformly larger than the sample variances of the MLE. Our study on mixtures of geometric distributions yields similar results. By comparing the estimation errors in Tables 4.5 and 4.6 to the ones in Tables 4.2, 4.3 and 4.4, we instantly know that the variance of the moment estimator of $\hat{b}$ is larger than the one provided by the MLE to a large extent. We have mentioned before that there is no guarantee that the estimates of a and b are in the interval $(0, 1)$; they may be negative, imaginary or greater than one. This makes the moment estimator an unattractive candidate for the estimation problem of a mixture geometric.

Another reason for the poor estimation of b is the variation between the moments, as we discussed previously in Chapter 3. For a mixture of two geometric distributions with $r = 10$ and $p = 0.6$, the first component has a parameter $a = 0.1$ and the second component has a parameter $b = 0.6513$. Let us denote z_{ak} as the theoretical k^{th} *normalised* moment of the first geometric component, z_{bk} as the the theoretical k^{th} *normalised* moment of the second geometric component, and z_k as the theoretical k^{th} *normalised* moment of the mixture

k	z_{ak}	z_{bk}	z_k	$\frac{z_{ak}}{z_{bk}}$
1	10	1.5353	6.6141	6.5132
2	100	2.3573	60.9430	42.4220
3	1000	3.6192	601.4477	276.3035

Table 4.7: Theoretical moments z_k of a sample with a mixture of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 0.6$, and the ratio of the moments of the two geometric components.

distribution; these theoretical moments are shown in Table 4.7. It is obvious that the effect of the second component on the third moment of the mixture is insignificant, compared to the first component. Therefore, to improve the estimation of b, we need a method that controls the variation between the moments.

4.3 The Method of Rising Factorial Fractional Moments

4.3.1 Introduction

In Chapter 3, we extended the traditional method of moments in an interesting way by replacing the integer k in moment μ_k with a fraction κ. Our simulation results confirm that the method of fractional moments improves the parameter estimation for a two-component exponential mixture model. Since the standard method of rising factorial moments provides estimates of b which are badly biased and have large variances, we shall now consider the modified approach, namely the method of rising factorial fractional moments, to solve the estimation problem for a geometric mixture. We desire a method that reduces the variation between moments so that the estimate of b becomes more promising.

We now study the method of rising factorial fractional moments on the discrete case, where the integer k is substituted by a fraction κ in (4.13) as follows:

$$\rho_\kappa = E\left[N\left(1+N\right)...\left(\kappa+N-1\right)\right] = E\left[\frac{\Gamma\left(N+\kappa\right)}{\Gamma\left(N\right)}\right]. \tag{4.17}$$

The latter expression in (4.17) is meaningful, as it allows κ to be, instead of an integer, any positive real number. Hence, we define the κ^{th} rising factorial fractional moment of N as

$$\rho_\kappa = E\left[\frac{\Gamma\left(N+\kappa\right)}{\Gamma\left(N\right)}\right]. \tag{4.18}$$

We next embark on finding the value of this moment:

$$\rho_\kappa = \sum_{n=1}^{\infty} \frac{\Gamma\left(n+\kappa\right)}{(n-1)!} f\left(n\right) = \sum_{n=0}^{\infty} \frac{\Gamma\left(n+1+\kappa\right)}{n!} \left[S\left(n\right) - S\left(n+1\right)\right]. \tag{4.19}$$

We can write (4.19) as

$$\rho_{\kappa}=\sum_{n=0}^{\infty}S\left(n\right)\left[\frac{\Gamma\left(n+1+\kappa\right)}{n!}-\frac{\Gamma\left(n+\kappa\right)}{\left(n-1\right)!}\right]=\sum_{n=0}^{\infty}\frac{\kappa\Gamma\left(n+\kappa\right)}{n!}S\left(n\right). \tag{4.20}$$

The sum in (4.20) can also be written as

$$\rho_{\kappa}=\Gamma\left(\kappa+1\right)\sum_{n=0}^{\infty}\frac{\kappa\left(\kappa+1\right)...\left(\kappa+n-1\right)}{n!}S\left(n\right). \tag{4.21}$$

For the discrete sojourn time N in a level of a hidden Markov chain with a transition matrix (4.1), the survival function is

$$S\left(n\right)=\frac{\boldsymbol{u}}{1-a}\boldsymbol{P}^{n}\mathbf{1}_{n}. \tag{4.22}$$

Putting (4.22) in (4.21) we get

$$\rho_{\kappa}=\frac{\Gamma\left(\kappa+1\right)\boldsymbol{u}}{1-a}\left(\boldsymbol{I}-\boldsymbol{P}\right)^{-\kappa}\mathbf{1}_{n}. \tag{4.23}$$

The *normalised* rising factorial moment is of course

$$z_{\kappa}=\frac{\rho_{\kappa}}{\Gamma\left(\kappa+1\right)}=\frac{\boldsymbol{u}}{1-a}\left(\boldsymbol{I}-\boldsymbol{P}\right)^{-\kappa}\mathbf{1}_{n}. \tag{4.24}$$

We know the distribution of N is a linear combination of two geometric distributions when there are two states hidden in the level. Therefore the survival function, with parameter vector $\boldsymbol{\Theta}=(a,b,p)$, is given by

$$S\left(n\right)=p\left(1-a\right)^{n}+\left(1-p\right)\left(1-b\right)^{n},$$

and hence the κ^{th} rising factorial moment is in the form of

$$\rho_{\kappa}=\Gamma\left(\kappa+1\right)\left[\frac{p}{a^{\kappa}}+\frac{\left(1-p\right)}{b^{\kappa}}\right]. \tag{4.25}$$

By comparing (4.25) with (3.78), we can see that even for fractional moments the first part of Theorem 2 holds.

Theorem 9 *(generalised) (Jalali (2005c))*

When between our continuous and discrete models the canonical relations we defined exists, then $\mu_{\kappa}=\rho_{\kappa}$, for all positive κ. (the rest of the theorem can similarly be generalised).

If $n_1,...,n_{n_o}$ is a sample of size n_o, we need to find the estimates of the *normalised* rising

κ	$z_{a\kappa}$	$z_{b\kappa}$	z_κ	$\frac{z_{a\kappa}}{z_{b\kappa}}$
0.1	1.2589	1.0438	1.1729	1.2061
0.2	1.5849	1.0895	1.3868	1.4547
0.3	1.9953	1.1373	1.6521	1.7544

Table 4.8: Theoretical moments z_κ of a sample with a mixture of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 0.6$, and the ratio of the moments of the two geometric components.

factorial fractional moments

$$\hat{z}_\kappa = \frac{1}{n_o \Gamma(\kappa+1)} \sum_{i=1}^{n_o} \frac{\Gamma(n_i+\kappa)}{\Gamma(n_i)} = \frac{1}{n_o \kappa} \sum_{i=1}^{n_o} \frac{\kappa(\kappa+1)...(\kappa+n_i-1)}{(n_i-1)!}. \tag{4.26}$$

The rest of the procedure is as in the analogous continuous case (from (3.80) to (3.86)). Since we need three moments, $\hat{z}_\kappa$, $\hat{z}_{\kappa_2}$ and $\hat{z}_{\kappa_3}$, we, like before, let $\kappa_2 = 2\kappa$ and $\kappa_3 = 3\kappa$ for simplicity.

We shall now demonstrate how a modification made on the standard method of rising factorial moments helps in controlling the variation between moments for a mixture of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 0.6$. We denote $z_{a\kappa}$ as the theoretical κ^{th} *normalised* moment of the first geometric component, $z_{b\kappa}$ as the the theoretical κ^{th} *normalised* moment of the second geometric component, and z_κ as the theoretical k^{th} *normalised* moment of the mixture distribution; the theoretical moments (when $\kappa = 0.1$) are shown in Table 4.8. Compare these with the ones given by the standard approach in Table 4.7, the ratio $\frac{z_{a\kappa}}{z_{b\kappa}}$ is significantly smaller; when $k = 1$, $\frac{z_{a3}}{z_{b3}}$ is 276 and this ratio is greatly reduced to 1.7544 when $\kappa = 0.1$ is used. Both components now have a similar effect on the mixture moments, so the fractional moment estimator should be able to provide more reasonable estimates of b, compared to the ordinary moment estimator.

4.3.2 Simulation Results

Now, let us study the simulation results of the method of rising factorial fractional moments from Tables 4.9 to 4.11. Like before, we consider three degrees of separation: $r = (2, 5, 10)$ for samples of different sizes $n_o = (10, 15, 20, 50, 1000)$. We took ten values of κ ranging from 0.1 to 1 with an increment of 0.1. For each κ, we generated 10000 data sets, each consisting of n_o observations, arising from a two-component geometric mixture model with $a = 0.1$, $b = 1 - 0.9^r$ and $p = 0.6$. We then fitted every data sets with three observed *normalised* rising factorial fractional moments $\hat{z}_\kappa$, $\hat{z}_{2\kappa}$ and $\hat{z}_{3\kappa}$. The measures of errors, namely the bias^2, variances and mean square errors, of the resulting 10000 estimators $\hat{a}$, $\hat{b}$ and $\hat{p}$ were calculated and we present the minimum of these errors in the tables. The corresponding κ's are shown in brackets in the tables.

Looking at Table 4.9, it is obvious that, for any sample size n_o, we should use a large fraction ($\kappa \geq 0.8$) to estimate a and b for a geometric mixture with a small separation $r = 2$.

$r=2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}}-a)^2$	8.15×10^{-7} (0.4)	6.60×10^{-8} (0.3)	3.24×10^{-8} (0.5)	5.87×10^{-6} (0.1)	1.17×10^{-8} (1)
$(\bar{\hat{b}}-b)^2$	3 (0.9)	0.1380 (1)	0.1178 (1)	3 (0.9)	0.0262 (1)
$(\bar{\hat{p}}-p)^2$	1.85×10^{-5} (0.8)	4.39×10^{-5} (0.7)	0.0002 (0.6)	0.0002 (0.4)	2.63×10^{-5} (0.9)
$Var[\hat{a}]$	0.0032 (1)	0.0025 (1)	0.0020 (1)	0.0014 (1)	0.0003 (0.9)
$Var\left[\hat{b}\right]$	767 (0.9)	722 (0.8)	150 (1)	232 (0.8)	8 (0.8)
$Var[\hat{p}]$	0.4403 (0.1)	0.3450 (0.1)	0.3108 (1)	0.2269 (0.1)	0.0708 (0.7)
$MSE[\hat{a}]$	0.0032 (1)	0.0025 (1)	0.0020 (1)	0.0014 (1)	0.0003 (0.9)
$MSE\left[\hat{b}\right]$	770 (0.9)	725 (0.8)	151 (1)	5703 (0.9)	8 (0.8)
$MSE[\hat{p}]$	0.4935 (0.1)	0.3827 (0.1)	0.3164 (1)	0.2438 (0.1)	0.0710 (0.7)

Table 4.9: Performance of the method of rising factorial fractional moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 0.6$ for different sample size n_o.

For p, we should use $\kappa = 0.1$ for small samples ($n_o \leq 50$) and $\kappa = 0.7$ for large samples ($n_o = 1000$). By now, we should not be surprised to see large $Var\left[\hat{b}\right]$ for samples with a small separation $r = 2$ and small sample size. Our previous study on the continuous analogue showed that the existence of a few large estimates of b makes $Var\left[\hat{b}\right]$ extremely large. We learned how a small y in (3.86) causes an over-estimation of b. Let us know focus on Figure 4.7 and study why $Var\left[\hat{b}\right]$ is 8 when $n_o = 1000$ in Table 4.9. The figure shows the scatter plot of $\hat{b}$ and $\hat{y}$; we can see that there exists six $\hat{b}$'s with values greater than 50. If we exclude these six outliers from the 10000 estimates of b, the variance of $\hat{b}$ is then reduced to 2. We also note that the corresponding values of $\hat{y}$ of these outliers are very small and close to 0.

Since the agreement between small samples and large samples on the best κ which minimises the variances of estimator b is not good for $r = 5$ and $r = 10$, as seen in Tables 4.10 and 4.11, we shall first analyse the large sample results for these two cases. For geometric mixtures with $r = 5$, the best fraction for estimating all three parameters is undoubtedly $\kappa = 0.2$, in terms of both the variance and the mean square error. When the components are better separated with $r = 10$, the best fraction for all parameters reduces to $\kappa = 0.1$. In other words, the best fraction reduces with r.

In both tables, we found that the best fraction for b appears to be $\kappa = 1$ when samples are small in sizes. This made us think that the ordinary moments perform better in small

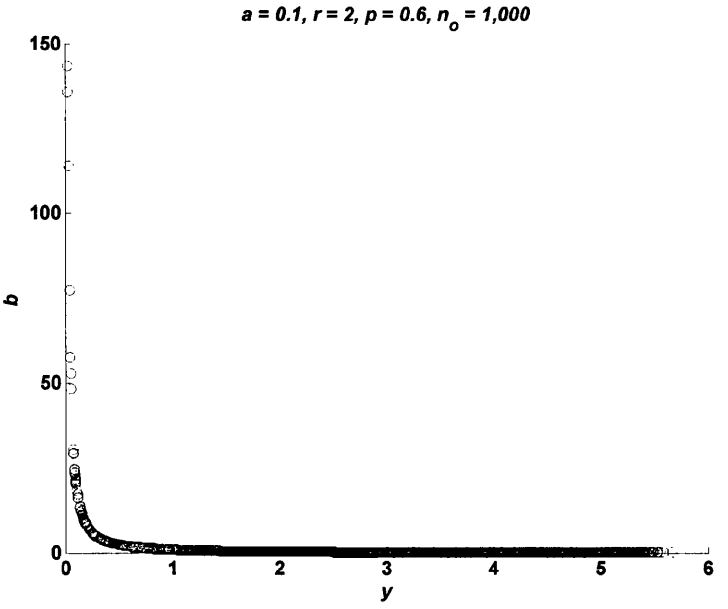


Figure 4.7: Plot of $\hat{b}$ versus $\hat{y}$ when $\kappa = 0.8$ is used to estimate from 10000 data sets, each consisting of 1000 observations, arising from a mixture of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 0.6$.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	1.27×10^{-8} (0.9)	2.41×10^{-6} (0.8)	4.15×10^{-7} (0.8)	9.29×10^{-8} (0.3)	3.51×10^{-9} (0.2)
$(\hat{b} - b)^2$	0.3522 (1)	0.1231 (1)	0.0602 (1)	0.1145 (1)	1.11×10^{-5} (0.1)
$(\hat{p} - p)^2$	0.0014 (0.4)	0.0002 (0.6)	3.83×10^{-6} (0.3)	1.05×10^{-5} (0.8)	4.80×10^{-7} (0.3)
$Var[\hat{a}]$	0.0048 (1)	0.0032 (1)	0.0023 (0.6)	0.0010 (0.6)	4.32×10^{-5} (0.2)
$Var[\hat{b}]$	3204 (1)	290 (1)	2098 (1)	1136 (1)	0.0025 (0.2)
$Var[\hat{p}]$	0.2400 (0.3)	0.1570 (0.1)	0.1236 (0.4)	0.0541 (0.7)	0.0032 (0.2)
$MSE[\hat{a}]$	0.0053 (1)	0.0032 (0.8)	0.0023 (0.6)	0.0010 (0.8)	4.32×10^{-5} (0.2)
$MSE[\hat{b}]$	3204 (1)	290 (1)	2098 (1)	1136 (1)	0.0025 (0.2)
$MSE[\hat{p}]$	0.2421 (0.4)	0.1588 (0.1)	0.1238 (0.4)	0.0543 (0.7)	0.0032 (0.2)

Table 4.10: Performance of the method of rising factorial fractional moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	0.0002 (0.9)	2.14×10^{-5} (0.9)	1.16×10^{-5} (0.9)	5.48×10^{-6} (0.9)	1.73×10^{-8} (0.1)
$(\bar{\hat{b}} - b)^2$	0.2014 (1)	0.2548 (1)	3.6516 (1)	0.8165 (0.2)	4.98×10^{-6} (0.1)
$(\bar{\hat{p}} - p)^2$	0.0094 (0.5)	4.30×10^{-6} (0.6)	3.73×10^{-5} (0.1)	4.68×10^{-6} (0.1)	2.78×10^{-8} (0.1)
$Var[\hat{a}]$	0.0057 (0.5)	0.0033 (0.4)	0.0021 (0.3)	0.0006 (0.1)	2.52×10^{-5} (0.1)
$Var[\hat{b}]$	1062 (1)	2687 (1)	24667 (1)	5946 (1)	0.0020 (0.1)
$Var[\hat{p}]$	0.1064 (0.5)	0.0731 (0.4)	0.0506 (0.4)	0.0213 (0.1)	0.0010 (0.1)
$MSE[\hat{a}]$	0.0063 (0.3)	0.0035 (0.1)	0.0023 (0.3)	0.0006 (0.1)	2.53×10^{-5} (0.1)
$MSE[\hat{b}]$	1062 (1)	2688 (1)	24671 (1)	5947 (1)	0.0020 (0.1)
$MSE[\hat{p}]$	0.1158 (0.5)	0.0746 (0.4)	0.0521 (0.4)	0.0213 (0.1)	0.0010 (0.1)

Table 4.11: Performance of the method of rising factorial fractional moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 0.6$ for different sample size n_o.

samples. We investigated the 10000 estimators and found that most of the fitted distributions are unreasonable when $\kappa = 1$. For instance, for $r = 5$ with $n_o = 10$, 70.05% of the estimates of b are negative; whereas for $r = 10$ with the same sample size, we observed 78.58% negative $\hat{b}$'s. Since b is a probability, it is impossible to be negative and so we do not recommend the use of $\kappa = 1$ for the estimation of b.

We found that the best κ for $r = 10$ is 0.1 when the number of observations in a sample is 1000. When $n_o = 10$, the resulted variance of $\hat{b}$ is extremely large although most of the fitted distributions given by $\kappa = 0.1$ are more reasonable compared to $\kappa = 1$. The reason for the large variance is 3.22% of $\hat{b}$'s are severely over-estimated and have values greater than 100. The maximum $\hat{b}$ is 2.20×10^{30}, which is extremely large because of a poor estimate $\hat{y} = 0.0009$.

We can confirm that the fractional moment estimator performs better than the standard moment estimator by comparing the measures of errors in Table 4.9 to 4.11 with the ones in Tables 4.5 and 4.6. For example, the estimation of the parameters in a geometric mixture with $r = 5$ and $n_o = 1000$ is greatly improved when $\kappa = 0.2$ is used; $Var[\hat{a}]$ is reduced from 0.0001 to 4.32×10^{-5}, $Var[\hat{b}]$ is significantly reduced from 166 to 0.0025 and $Var[\hat{p}]$ is decreased from 0.0156 to 0.0032. We also note that the performance of this method is greatly enhanced by each of the following factors: larger sample size and better separation.

4.3.3 Asymptotic Covariance Matrix of the Rising Factorial Fractional Moment Estimators

In this section, we find an approximation for the theoretical asymptotic variance of rising factorial fractional moment estimator for a mixture of two geometric distributions. From (4.25), the κ^{th} rising factorial moment is given by

$$\rho_\kappa = \Gamma(\kappa+1)\left[\frac{p}{a^\kappa} + \frac{(1-p)}{b^\kappa}\right].$$

We follow the procedure in Section 1.5.5 to find an approximation for the covariance matrix $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ of the rising factorial fractional moment estimator $\boldsymbol{\Theta} = (a, b, p)$, which is given by

$$\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right] \approx \boldsymbol{D}[\boldsymbol{\Theta}]^{-1}\boldsymbol{V}[\hat{\rho}]\left(\boldsymbol{D}[\boldsymbol{\Theta}]^{-1}\right)^T, \tag{4.27}$$

where $\boldsymbol{D}[\boldsymbol{\Theta}]$ is the Jacobian matrix with entries $d_{ij} = \dfrac{\partial \rho_{\kappa_i}}{\partial \boldsymbol{\Theta}_j}$ and $\boldsymbol{V}[\hat{\rho}]$ is the covariance matrix of the rising factorial fractional moments with entries $Cov[\hat{\rho}_\kappa, \hat{\rho}_\ell]$.

Covariance Matrix of the Rising Factorial Fractional Moments

The covariance matrix of the rising factorial fractional moments $\boldsymbol{V}[\hat{\rho}]$ has entries

$$Cov[\hat{\rho}_\kappa, \hat{\rho}_\ell] = \frac{1}{n_o}\left[E\left[(N)_\kappa (N)_\ell\right] - E\left[(N)_\kappa\right]E\left[(N)_\ell\right]\right], \tag{4.28}$$

where $(N)_\kappa$ represents the κ^{th} rising factorial moment. We denote $\xi(\kappa, \ell \mid a)$ as the expectation of the product of $(N)_\kappa$ and $(N)_\ell$ for a single geometric distribution with probability a. This expectation is defined as

$$\begin{aligned}\xi(\kappa, \ell \mid a) &= E\left[\frac{\Gamma(N+\kappa)\Gamma(N+\ell)}{\Gamma(N)^2}\right] \\ &= a\Gamma(\kappa+1)\Gamma(\ell+1)\sum_{n=1}^{\infty}(1-a)^{n-1}\frac{(\kappa+1)_{n-1}(\ell+1)_{n-1}}{(1)_{n-1}(n-1)!}.\end{aligned} \tag{4.29}$$

The power series involved in (4.29) is just the Gaussian Hypergeometric function with parameters $(\kappa, \ell; 1)$ (see Slater (1966)). Hence, (4.29) is given by

$$\xi(\kappa, \ell \mid a) = a\Gamma(\kappa+1)\Gamma(\ell+1)_2F_1(\kappa+1, \ell+1; 1 \mid 1-a). \tag{4.30}$$

It is well-known (the so-called Kummer's solutions) that the Hypergeometric function in (4.30) can be expressed as

$${}_2F_1(\kappa+1, \ell+1; 1 \mid 1-a) = a^{-\kappa-\ell-1}{}_2F_1(-\kappa, -\ell; 1 \mid 1-a). \tag{4.31}$$

As the ${}_2F_1(-\kappa,-\ell;1\mid 1-a)$ series converges, we prefer this representation. Note that when κ and ℓ are positive integers, the series has only a finite number of terms, and we have a polynomial in a. So, we express $\xi(\kappa,\ell\mid a)$ as

$$\xi(\kappa,\ell\mid a)=a^{-\kappa-\ell}\Gamma(\kappa+1)\,\Gamma(\ell+1)\,{}_2F_1(-\kappa,-\ell;1\mid 1-a). \tag{4.32}$$

Using (4.32), we find the expectation of the product of $(N)_\kappa$ and $(N)_\ell$ for a mixture of two geometric distributions with parameter vector $\boldsymbol{\Theta}=(a,b,p)$ as follows:

$$\begin{aligned} E[(N)_\kappa(N)_\ell] &= pE[(N)_\kappa(N)_\ell\mid a]+(1-p)\,E[(N)_\kappa(N)_\ell\mid b] \\ &= \Gamma(\kappa+1)\,\Gamma(\ell+1)\left[\begin{array}{c} pa^{-\kappa-\ell}{}_2F_1(-\kappa,-\ell;1\mid 1-a) \\ +(1-p)\,b_2^{-\kappa-\ell}F_1(-\kappa,-\ell;1\mid 1-b)\end{array}\right]. \end{aligned} \tag{4.33}$$

Hence, substituting (4.33) and (4.25) into (4.28) we obtain the covariance between $\hat{\rho}_\kappa$ and $\hat{\rho}_\ell$ as

$$Cov[\hat{\rho}_\kappa,\hat{\rho}_\ell]=\frac{1}{n_o}\Gamma(\kappa+1)\,\Gamma(\ell+1)\left[\begin{array}{c} pa^{-\kappa-\ell}{}_2F_1(-\kappa,-\ell;1\mid 1-a) \\ +(1-p)\,b^{-\kappa-\ell}{}_2F_1(-\kappa,-\ell;1\mid 1-b) \\ -\left(\begin{array}{c} p^2a^{-\kappa-\ell} \\ +p(1-p)\left(a^{-\kappa}b^{-\ell}+a^{-\ell}b^{-\kappa}\right) \\ +(1-p)^2\,b^{-\kappa-\ell}\end{array}\right)\end{array}\right]. \tag{4.34}$$

Hence, the ij^{th} entry of the covariance matrix of the rising factorial fractional moments $\boldsymbol{V}[\hat{\rho}]$ is given by $Cov[\hat{\rho}_i,\hat{\rho}_j]$ according to the expression in (4.34).

Jacobian Matrix of the Rising Factorial Fractional Moment Estimator

Let $\boldsymbol{D}[\boldsymbol{\Theta}]$ be the Jacobian Matrix with entries: $d_{ij}=\dfrac{\partial\rho_{\kappa_i}}{\partial\boldsymbol{\Theta}_j}$,

$$\boldsymbol{D}[\boldsymbol{\Theta}]=\left[\begin{array}{ccc} \frac{\partial\rho_{\kappa_1}}{\partial a} & \frac{\partial\rho_{\kappa_1}}{\partial b} & \frac{\partial\rho_{\kappa_1}}{\partial p} \\ \frac{\partial\rho_{\kappa_2}}{\partial a} & \frac{\partial\rho_{\kappa_2}}{\partial b} & \frac{\partial\rho_{\kappa_2}}{\partial p} \\ \frac{\partial\rho_{\kappa_3}}{\partial a} & \frac{\partial\rho_{\kappa_3}}{\partial b} & \frac{\partial\rho_{\kappa_3}}{\partial p}\end{array}\right], \tag{4.35}$$

where

$$\frac{\partial\rho_{\kappa_i}}{\partial a}=-p\Gamma(\kappa_i+1)\,\kappa_i a^{-(\kappa_i+1)}, \tag{4.36}$$

$$\frac{\partial\rho_{\kappa_i}}{\partial b}=-(1-p)\,\Gamma(\kappa_i+1)\,\kappa_i b^{-(\kappa_i+1)}, \tag{4.37}$$

and

$$\frac{\partial\rho_{\kappa_i}}{\partial p}=\Gamma(\kappa_i+1)(a^{-\kappa_i}-b^{-\kappa_i}) \tag{4.38}$$

r	Optimal κ	$Var[\hat{a}]$
2	0.5348	0.0003
3	0.3532	9.22×10^{-5}
4	0.2502	5.64×10^{-5}
5	0.1788	4.26×10^{-5}
6	0.1243	3.56×10^{-5}
7	0.0857	3.13×10^{-5}
8	0.0405	2.84×10^{-5}
9	0.0100	2.64×10^{-5}
10	0.0100	2.50×10^{-5}

Table 4.12: Optimal fraction κ and theoretical minimum variance of the rising factorial fractional moment estimator $\hat{a}$ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$.

for $i = 1, 2, 3$. For simplicity, we set κ_2 as 2κ and κ_3 and 3κ for our investigation.

Approximated Asymptotic Covariance Matrix of the Rising Factorial Fractional Moment Estimator

Now, we can find the approximated covariance matrix of the rising factorial fractional moment estimator by substituting the covariance matrix of the rising factorial fractional moment $\boldsymbol{V}[\hat{p}]$ using (4.34) and the Jacobian matrix $\boldsymbol{D}[\boldsymbol{\Theta}]$ from (4.35) into (4.27). By substituting the true values of $\boldsymbol{\Theta} = (a, b, p)$ and n_o into (4.27), we obtain an approximation for the theoretical variance of the rising factorial fractional moment estimator $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ for a mixture of two geometric distributions.

4.3.4 Optimal κ

The diagonal entries of $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ given by (4.27) are $Var[\hat{a}]$, $Var\left[\hat{b}\right]$ and $Var[\hat{p}]$ which are all in terms of a, b, p, κ and n_o. These expressions allow us to make use of the built-in function "FindMinimum" in Mathematica to find the values of κ's which minimise the variances of the estimators in theory. We consider mixtures of two geometric distributions with true parameters $a = 0.1$, $b = 1 - 0.9^r$ and $p = 0.6$, with nine different degrees of separation between the two components, $r = (2, 3, ..., 10)$.

Figure 4.8 shows nine plots of $Var[\hat{a}]$ versus κ in which each plot represents different r; whereas Table 4.12 presents the optimal value of κ for estimating a and the corresponding minimum variance for each r. Theoretically, the best κ for a when $r = 2$ is 0.5348 with minimum $Var[\hat{a}] = 0.0003$; the best κ for a decreases with r, as seen in both Table 4.12 and Figure 4.8. Indeed, the best κ is below 0.1 for $r \geq 7$. We can see that the plots for $r \geq 7$ are quite flat near to the origin; this means that any κ less than 0.1 should be able to provide low variance of $\hat{a}$ for mixture populations that are well separated.

In Figure 4.9, we show nine plots of $Var\left[\hat{b}\right]$ versus κ; whereas Table 4.13 outlines the optimal κ for estimating b and the resulting minimum variance for each r. The optimal κ's

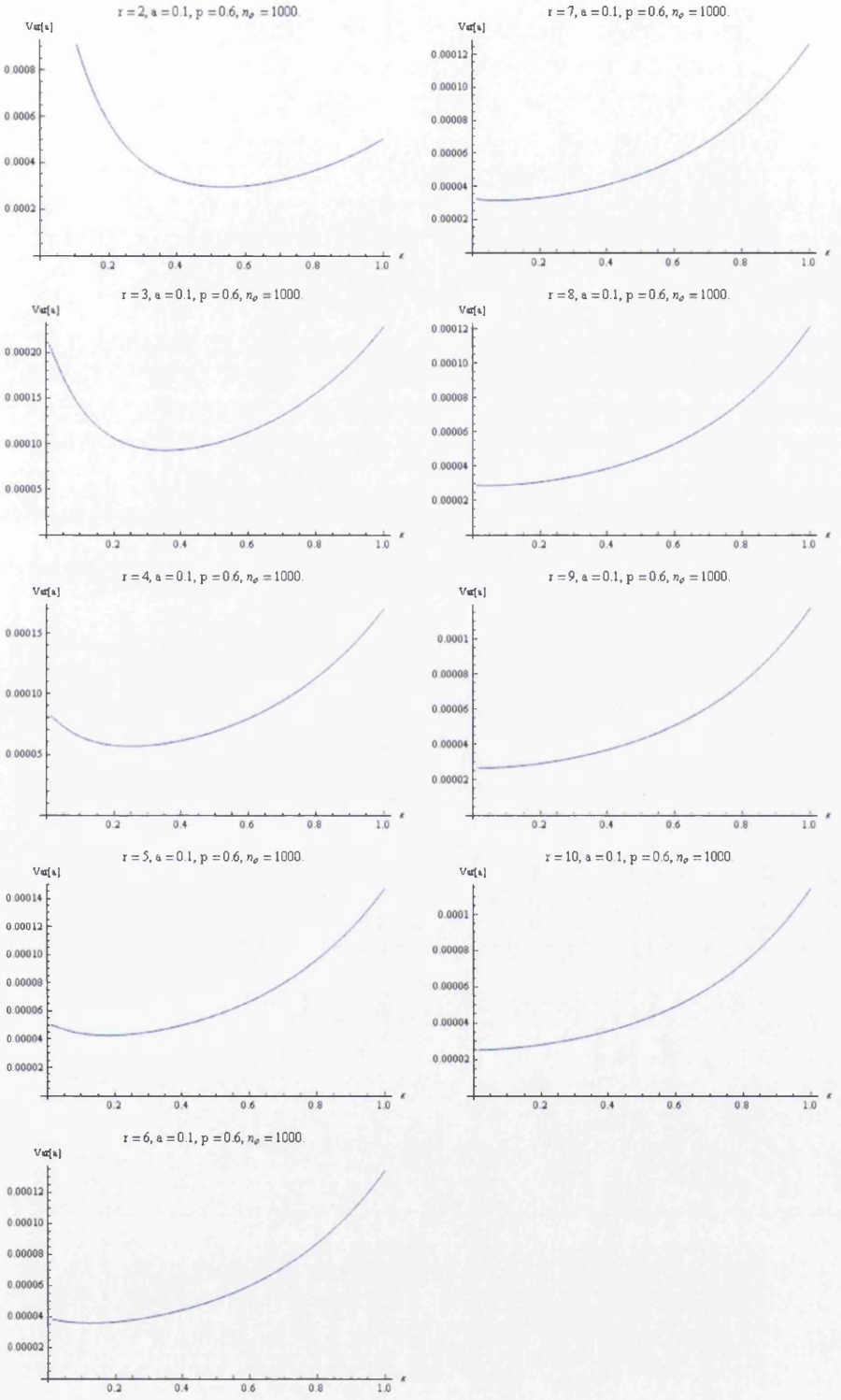


Figure 4.8: Plots of $Var[\hat{a}]$ versus κ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$.

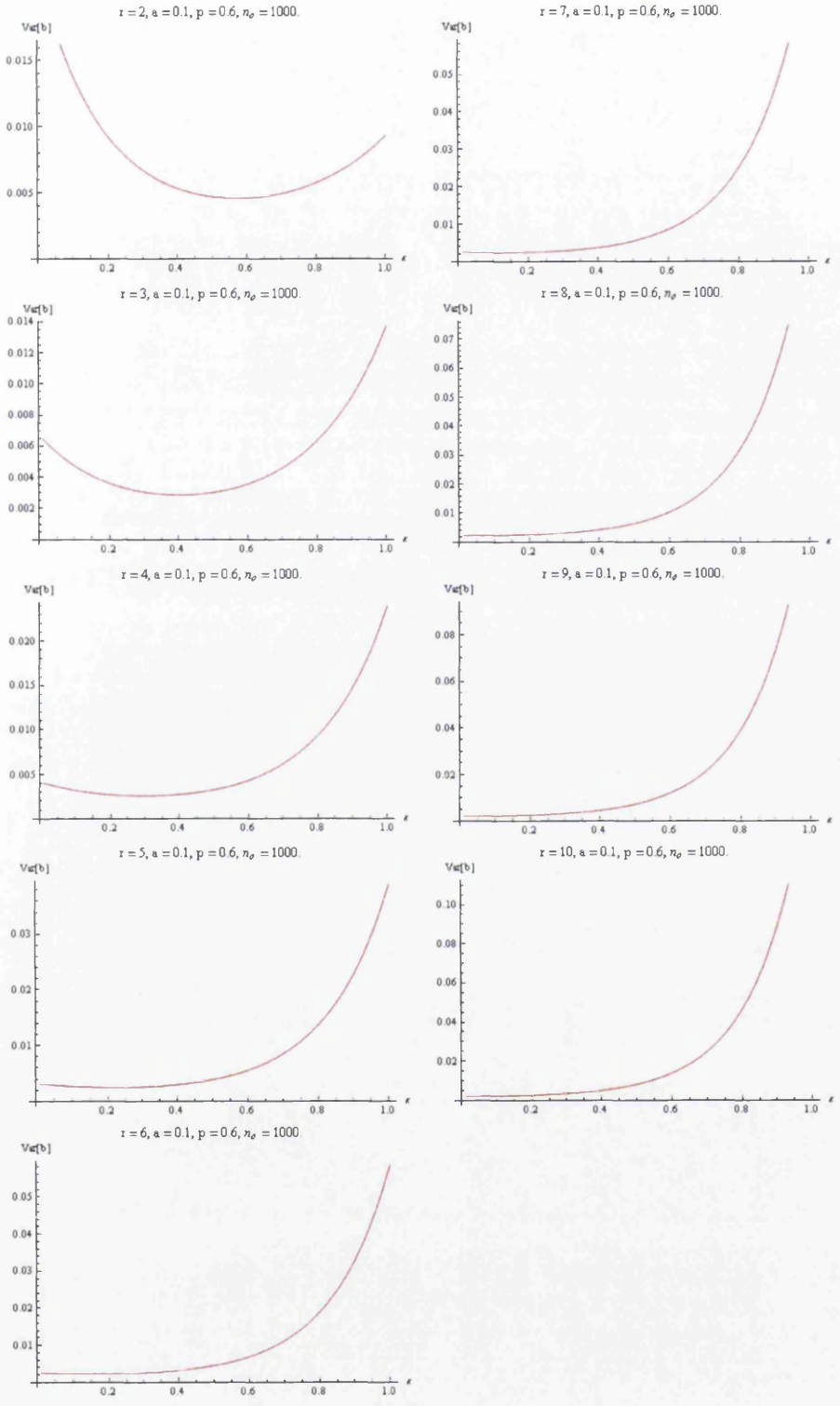


Figure 4.9: Plots of $Var\left[\hat{b}\right]$ versus κ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$.

r	Optimal κ	$Var\left[\hat{b}\right]$
2	0.5720	0.0045
3	0.4019	0.0028
4	0.3002	0.0025
5	0.2259	0.0024
6	0.1658	0.0023
7	0.1141	0.0022
8	0.0700	0.0021
9	0.0220	0.0019
10	0.0154	0.0018

Table 4.13: Optimal fraction κ and theoretical minimum variance of the rising factorial fractional moment estimator $\hat{b}$ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$.

r	Optimal κ	$Var\left[\hat{p}\right]$
2	0.5512	0.1288
3	0.3714	0.0156
4	0.2648	0.0058
5	0.1875	0.0032
6	0.1257	0.0021
7	0.0732	0.0016
8	0.0512	0.0013
9	0.0100	0.0011
10	0.0100	0.0009

Table 4.14: Optimal fraction κ and theoretical minimum variance of the rising factorial fractional moment estimator $\hat{p}$ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$.

for b are slightly larger than the ones for a. For mixtures with little separation between the two components, the optimal κ for b is larger; for instance, $Var\left[\hat{b}\right]$ for $r = 2$ is minimised when $\kappa = 0.5720$. Similarly to a, as shown in Figure 4.13, the plots are flat for $\kappa \leq 0.2$ when $r \geq 6$. In other words, for mixtures which are well separated, any $\kappa \leq 0.2$ should be able to provide estimates of b with low variance.

For p, we show plots of $Var\left[\hat{p}\right]$ versus κ in Figure 4.10 and the minimum variance of $\hat{p}$ for each r with its counterpart κ in Table 4.14. The trend is similar to the other two parameters: the optimal κ decreases with r. For mixtures with $r \geq 7$, we can use any $\kappa \leq 0.1$ to enhance the precision of $\hat{p}$.

We shall now validate the conformity between theory and practice with some simulations. For each of the nine r's, we simulated 10000 data sets arising from the specified mixture geometric distributions and estimated every sample with ten κ's ranging from 0.1 to 1 with an increment of 0.1. We found the minimum variances of the estimators a, b and p and presented them along with their corresponding κ in Tables 4.15, 4.16 and 4.17. For simplicity, we round up the theoretical values of κ and the variances of the estimators shown

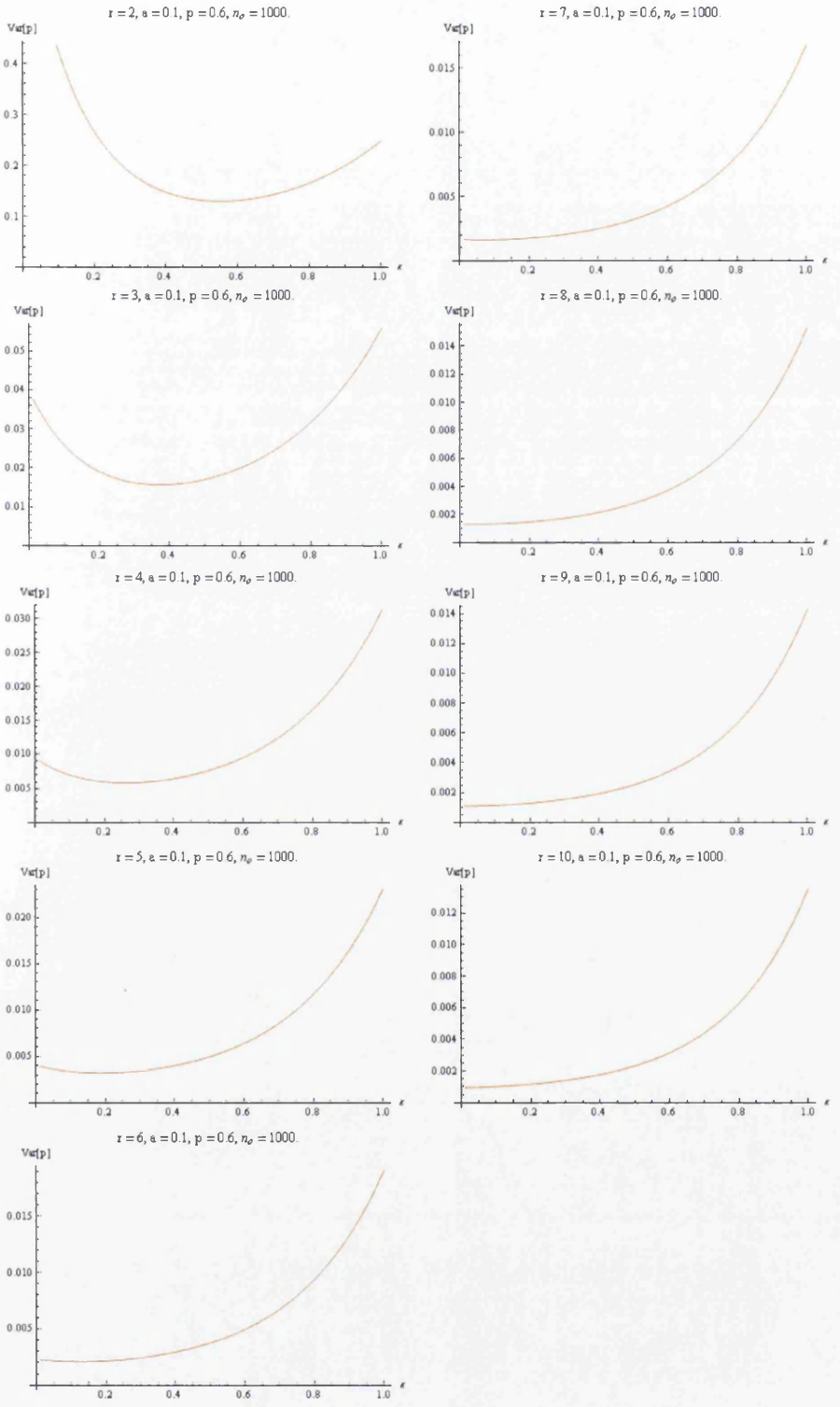


Figure 4.10: Plots of $Var[\hat{p}]$ versus κ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$.

r	Theoretical Optimal κ	Practical Optimal κ	Theoretical $Var[\hat{a}]$	Practical $Var[\hat{a}]$
2	0.5	0.9	0.0003	0.0003
3	0.4	0.4	9.31×10^{-5}	9.66×10^{-5}
4	0.3	0.3	5.70×10^{-5}	5.85×10^{-5}
5	0.2	0.2	4.27×10^{-5}	4.32×10^{-5}
6	0.2	0.1	3.56×10^{-5}	3.58×10^{-5}
7	0.1	0.1	3.13×10^{-5}	3.18×10^{-5}
8	0.1	0.1	2.88×10^{-5}	2.94×10^{-5}
9	0.1	0.1	2.71×10^{-5}	2.76×10^{-5}
10	0.1	0.1	2.59×10^{-5}	2.52×10^{-5}

Table 4.15: Theoretical and simulated minimum variance of rising factorial fractional moment estimator $\hat{a}$ given by the optimal κ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical Optimal κ	Practical Optimal κ	Theoretical $Var[\hat{b}]$	Practical $Var[\hat{b}]$
2	0.6	0.8	0.0045	8
3	0.4	0.3	0.0028	0.0037
4	0.3	0.2	0.0025	0.0028
5	0.2	0.2	0.0024	0.0025
6	0.2	0.1	0.0023	0.0023
7	0.1	0.1	0.0022	0.0022
8	0.1	0.1	0.0021	0.0021
9	0.1	0.1	0.0020	0.0021
10	0.1	0.1	0.0020	0.0020

Table 4.16: Theoretical and simulated minimum variance of rising factorial fractional moment estimator $\hat{b}$ given by the optimal κ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical Optimal κ	Practical Optimal κ	Theoretical $Var[p]$	Practical $Var[\hat{p}]$
2	0.6	0.7	0.1303	0.0708
3	0.4	0.4	0.0156	0.0156
4	0.3	0.3	0.0058	0.0059
5	0.2	0.2	0.0032	0.0032
6	0.1	0.1	0.0021	0.0022
7	0.1	0.1	0.0016	0.0016
8	0.1	0.1	0.0013	0.0013
9	0.1	0.1	0.0011	0.0011
10	0.1	0.1	0.0010	0.0010

Table 4.17: Theoretical and simulated minimum variance of rising factorial fractional moment estimator $\hat{p}$ given by the optimal κ for a mixture of two geometric distributions with various r and fixed $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

in Tables 4.12 to 4.14. Since the theoretical best κ for large r (≥ 7) are generally lower than 0.1 and we know that the plots are flat for $\kappa \leq 0.1$, we round up the κ to 0.1 for large r. As seen in these tables, both the best κ's and the observed variances have strong agreement with the theory, except for $r = 2$. To conclude, we should use a large κ to estimate a two-component geometric distribution with little separation between populations. As the components become further and further from each other, we should use a smaller κ; for $r \geq 7$, any κ less than 0.1 should be able to provide good parameter estimates.

In practice, we do not know the separation between the components, how should we choose a good κ to ensure the precision of the estimates? This issue has been discussed and solved in Chapter 3 for the case of the exponential mixture. Surely, a similar approach can be adopted for the geometric mixture. Given a data set, we should estimate the parameters with, say, ten different values of κ, ranging from 0.1 to 1. We then substitute the yielded sets of estimates into (4.27) and choose the set of estimates with the minimum $Var\left[\hat{\Theta}\right]$. To prove that we can do this, we simulated a data set, consisting of 1000 observations, from a mixture of two geometric distributions with true parameters $a = 0.1$, $b = 0.4095$ and $p = 0.6$, and estimated the parameters with ten different κ. After substituting the estimates into (4.27), we plotted the resulted $Var\left[\hat{\Theta}\right]$ alongside the theoretical asymptotic $Var\left[\hat{\Theta}\right]$ (given by the true parameters) against κ, for $\hat{\Theta} = a$, b and p respectively in Figure 4.11. Undoubtedly, the agreement between the theory and practice is excellent; the estimates given by the optimal κ (in this case, the best κ for $r = 5$ is 0.2) do make $Var\left[\hat{\Theta}\right]$ smallest when they are put into (4.27).

4.3.5 Discussion

Like the ordinary moments of an exponential mixture, $Var[\rho_\kappa]$ increases with κ, as illustrated in Figure 4.12. For instance, when $\kappa = 1$, $Var[\rho_2]$ is 116 and $Var[\rho_3]$ is 358676

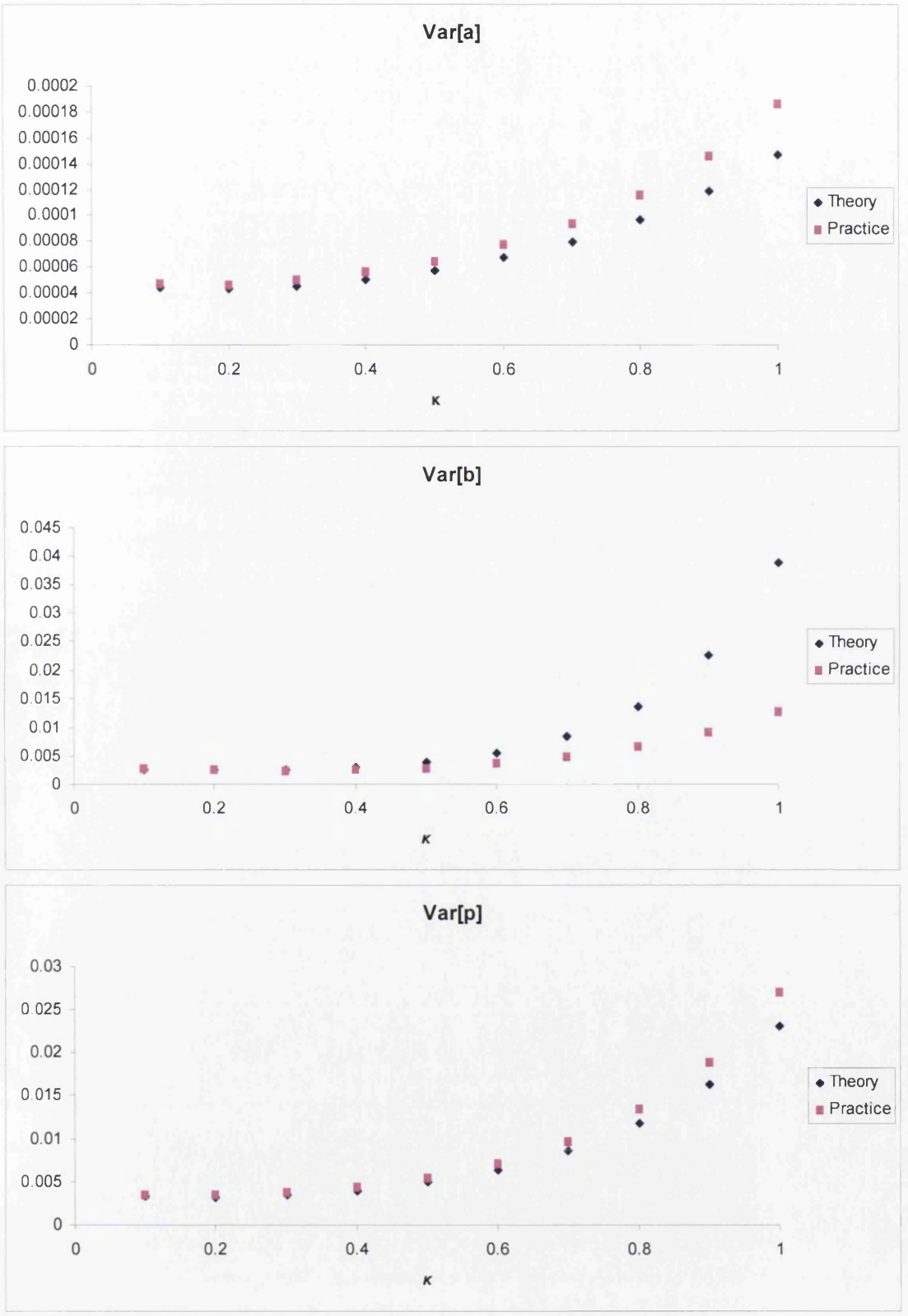


Figure 4.11: Asymptotic variance of the rising factorial fractional moment estimator given by true parameters and parameter estimates versus κ, based on a data set, consisting of 1000 observations, simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$.

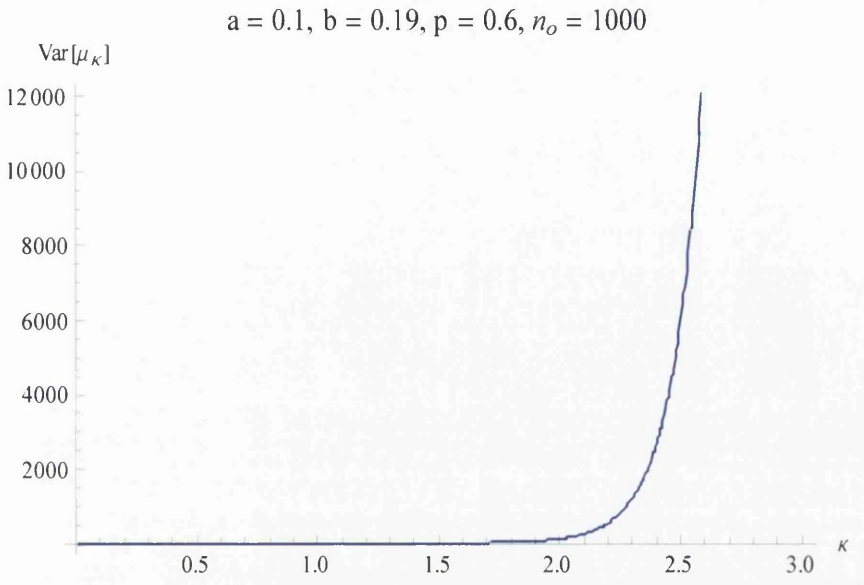


Figure 4.12: Plot of $Var[\rho_\kappa]$ versus κ for a mixture of two geometric distributions with $a = 0.1$, $b = 0.19$, $p = 0.6$ and $n_o = 1000$.

for a mixture of two geometric distributions with $a = 0.1$, $r = 2$, $p = 0.6$ and $n_o = 1000$. These large variances of the moments result in poor estimates of the parameters. On the other hand, when $\kappa = 0.6$ is used, $Var[\rho_2]$ and $Var[\rho_3]$ are greatly reduced to, respectively, 0.2913 and 25. Therefore, a smaller κ, with a lower variance of the moment, yields more accurate estimates.

With the use of fractional moments, the efficiency of the usual method of rising factorial moments are undoubtedly greatly increased for a mixture of two geometric distributions. However, the potential problems for moment based methods still apply for this estimator. The estimation of b, once again, appear to be more difficult compared to a and p. The fractional moment estimator is quite likely to over-estimate b, particularly for small samples. As we mentioned before, the reason for this is the bad estimation of y; as seen in Figure 3.17, when $\hat{y}$ is smaller than its true value and close to the origin, the estimate of b can be extremely large. Additionally, the estimates of all three parameters may not automatically lie in the interval $(0, 1)$; they may be imaginary or greater than 1. Nevertheless, the probability of obtaining such bad estimates decreases with sample size.

This method becomes more robust as the degree of separation increases, especially for samples of large sizes. We found an approximation for the theoretical asymptotic variances of the estimators which has a sound agreement with simulations. Hence, it allows us to suggest good fractions for parameter estimation of mixtures with different degrees of separation between populations. For estimating the parameters in a mixture where the two components are not well separated, we should use a large κ; For mixtures with $r \geq 6$, we can use $\kappa = 0.1$ or any fraction lower than this value to obtain highly precise estimates. We have also

suggested a solution for users to choose the best κ when they do not know the separation between the components in practice. One should simply estimate the parameters with a few different κ and choose the set of parameters which minimises $Var\left[\hat{\Theta}\right]$ when they are put into (4.27).

For our investigation, we fixed the second fractional moment z_{κ_2} as $z_{2\kappa}$ and the third fractional moment z_{κ_3} as $z_{3\kappa}$, greater efficiency might be achieved if the values of these two fractions (κ_2 and κ_3) are allowed to differ.

4.4 The Method of Attenuated Rising Factorial Fractional Moments

4.4.1 Introduction

As seen in the previous chapter, spectacular gains in efficiency can be obtained if we make a small attenuation on the fractional moments used to estimate the parameters in a mixture of two exponential distributions. We shall now consider a similar approach for the discrete analogue by modifying the method of rising factorial fractional moments described in the previous section. Let us consider the expectation $E\left[\frac{\Gamma(N+\kappa)}{\Gamma(N)}\exp(-cN)\right]$ and denote it as $\rho_\kappa(c)$. Therefore, for a mixture of two geometric distributions with PMF in the form of (4.4), the c-attenuated rising factorial moment of order κ is given by

$$\begin{aligned}\rho_\kappa(c) &= \sum_{n=1}^{\infty}\varepsilon^n\frac{\Gamma(n+\kappa)}{\Gamma(n)}\left(pa(1-a)^{n-1}+(1-p)b(1-b)^{n-1}\right) \\ &= \varepsilon\Gamma(\kappa+1)\left[\frac{pa}{(1-\varepsilon(1-a))^{\kappa+1}}+\frac{(1-p)b}{(1-\varepsilon(1-b))^{\kappa+1}}\right],\end{aligned} \tag{4.39}$$

where $\varepsilon=\exp(-c)$. We define the *normalised* c-attenuated rising factorial moment of order κ as

$$\begin{aligned}z_\kappa &= \frac{\rho_\kappa(c)}{\Gamma(\kappa+1)} \\ &= \varepsilon\left[\frac{pa}{(1-\varepsilon(1-a))^{\kappa+1}}+\frac{(1-p)b}{(1-\varepsilon(1-b))^{\kappa+1}}\right].\end{aligned}$$

For estimation purposes, we let

$$\delta_1 = pa\varepsilon(1-\varepsilon(1-a))^{-1}, \tag{4.40}$$

$$\delta_2 = (1-p)b\varepsilon(1-\varepsilon(1-b))^{-1},$$

$$x_1 = (1-\varepsilon(1-a))^{-\kappa},$$

$$x_2 = (1-\varepsilon(1-b))^{-\kappa},$$

and consider the following system of 4 $(= 2m)$ equations

$$\begin{aligned} z_0 &= \delta_1 + \delta_2, \\ z_\kappa &= \delta_1 x_1 + \delta_2 x_2, \\ z_{2\kappa} &= \delta_1 x_1^2 + \delta_2 x_2^2, \\ z_{3\kappa} &= \delta_1 x_1^3 + \delta_2 x_2^3. \end{aligned} \tag{4.41}$$

The solutions of x_1 and x_2 can be found by solving the m^{th} order determinantal equation

$$\det \begin{bmatrix} z_0 & z_\kappa & z_{2\kappa} \\ z_\kappa & z_{2\kappa} & z_{3\kappa} \\ 1 & u & u^2 \end{bmatrix} = 0.$$

In this case x_1 and x_2 are the roots of

$$u^2 - su + t = 0, \tag{4.42}$$

where

$$s = \frac{z_0 z_{3\kappa} - z_\kappa z_{2\kappa}}{z_0 z_{2\kappa} - z_\kappa^2} \tag{4.43}$$

and

$$t = \frac{z_\kappa z_{3\kappa} - z_{2\kappa}^2}{z_0 z_{2\kappa} - z_\kappa^2}. \tag{4.44}$$

Therefore,

$$x_1 = \frac{s + \sqrt{s^2 - 4t}}{2} \tag{4.45}$$

and

$$x_2 = \frac{s - \sqrt{s^2 - 4t}}{2}. \tag{4.46}$$

Having found x_1 and x_2, we shall proceed to find δ_1 and δ_2 from

$$\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = V^{-1} \begin{bmatrix} z_0 \\ z_\kappa \end{bmatrix}, \tag{4.47}$$

where V is the Vandermonde matrix based on x_1 and x_2:

$$V = \begin{bmatrix} 1 & 1 \\ x_1 & x_2 \end{bmatrix}. \tag{4.48}$$

Hence, from (4.47), we know

$$\delta_1 = \frac{z_0 x_2 - z_\kappa}{x_2 - x_1} \tag{4.49}$$

and

$$\delta_2 = \frac{z_\kappa - z_0 x_1}{x_2 - x_1}. \tag{4.50}$$

By now, we have found x_j's and δ_j's for $j = 1$ and 2, so we are able to find a and b from the following formulae

$$a = 1 - \varepsilon\left(1 - x_1^{-\frac{1}{\kappa}}\right) \tag{4.51}$$

and

$$b = 1 - \varepsilon\left(1 - x_2^{-\frac{1}{\kappa}}\right). \tag{4.52}$$

By substituting (4.51) and c into (4.40), the mixing weight p can be found from

$$p = \frac{\delta_1\left(1 - \varepsilon\left(1 - a\right)\right)}{a\varepsilon}. \tag{4.53}$$

If $n_1, ..., n_{n_o}$ is a sample of size n_o drawn from a two-component geometric mixture distribution, we obtain the attenuated rising factorial moment estimators by simply replacing z_κ with the sample moments

$$\hat{z}_\kappa = \frac{1}{n_o\Gamma\left(\kappa+1\right)}\sum_{i=1}^{n_o}\frac{\Gamma\left(n_i+\kappa\right)}{\Gamma\left(n_i\right)}\exp\left(-cn_i\right)$$

in (4.43) and (4.44), and following the estimation procedure from (4.45) to (4.53).

4.4.2 Simulation Results

In order to evaluate the robustness of using the method of attenuated rising factorial fractional moments for estimating the parameters in a mixture of two geometric distributions, a simulation experiment was carried out in which we considered 100 combinations of fraction κ and attenuation c: $\kappa = (0.1, ..., 1)$ and for each κ we take $c = (0.01, 0.02, ..., 0.1)$. One of our purposes is to investigate the ideal combinations (κ, c) for geometric mixtures with different degrees of separation between populations. For each combination, we simulated 10000 data sets, each of size n_o, drawn from a two component geometric mixture model with true parameters $a = 0.1$, $b = 1 - 0.9^r$ and $p = 0.6$. Like before, we considered five sample sizes $n_o = (10, 15, 20, 50, 1000)$ and three degrees of separation $r = (2, 5, 10)$. For each r and n_o, we found the lowest measures of errors out of the 100 values and recorded them in Tables 4.18, 4.19 and 4.20, in which the corresponding combinations of κ and c are presented in brackets.

Table 4.18 represents a situation with relatively little separation between the components ($r = 2$). It is clear from the estimation results that we should use a large fraction and a small attenuation to minimise the variances and mean square errors of all parameters for a mixture where the two components are not well separated. Although the variances of estimator $\hat{b}$ are still large for small samples, these values are indeed smaller than the ones obtained by the fractional moment estimator. Great improvement in the estimation of b is obtained when an attenuation is made on fractional moments, especially when the sample size is sufficiently large. Recall from Table 4.9 that $Var\left[\hat{b}\right]$ was 8 when $\kappa = 0.8$ was used on

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	6.25×10^{-8} $(0.5, 0.06)$	1.05×10^{-7} $(0.5, 0.1)$	5.29×10^{-12} $(0.2, 0.1)$	8.97×10^{-8} $(0.1, 0.07)$	2.89×10^{-6} $(1, 0.01)$
$(\bar{\hat{b}} - b)^2$	0.0044 $(1, 0.03)$	0.0061 $(1, 0.01)$	0.0012 $(1, 0.09)$	6.95×10^{-6} $(1, 0.06)$	0.0006 $(1, 0.06)$
$(\bar{\hat{p}} - p)^2$	1.93×10^{-10} $(0.7, 0.05)$	8.07×10^{-8} $(0.8, 0.01)$	3.49×10^{-8} $(0.9, 0.06)$	2.92×10^{-5} $(0.6, 0.03)$	1.41×10^{-6} $(0.5, 0.1)$
$Var[\hat{a}]$	0.0033 $(1, 0.01)$	0.0026 $(1, 0.01)$	0.0022 $(1, 0.01)$	0.0016 $(1, 0.01)$	0.0003 $(0.9, 0.01)$
$Var[\hat{b}]$	122 $(1, 0.01)$	122 $(1, 0.01)$	94 $(1, 0.02)$	96 $(1, 0.04)$	0.4134 $(0.7, 0.04)$
$Var[\hat{p}]$	0.4263 $(0.5, 0.02)$	0.4090 $(1, 0.02)$	0.2977 $(0.1, 0.01)$	0.2285 $(0.2, 0.01)$	0.0711 $(0.8, 0.01)$
$MSE[\hat{a}]$	0.0033 $(1, 0.01)$	0.0026 $(1, 0.01)$	0.0022 $(1, 0.01)$	0.0017 $(1, 0.01)$	0.0003 $(0.9, 0.01)$
$MSE[\hat{b}]$	122 $(1, 0.01)$	122 $(1, 0.01)$	94 $(1, 0.02)$	96 $(1, 0.04)$	0.8323 $(0.7, 0.04)$
$MSE[\hat{p}]$	0.4451 $(1, 0.04)$	0.4165 $(1, 0.02)$	0.3301 $(0.1, 0.01)$	0.2465 $(0.5, 0.03)$	0.0716 $(0.8, 0.01)$

Table 4.18: Performance of the method of attenuated rising factorial fractional moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 0.6$ for different sample size n_o.

samples of size 1000 with $r = 2$. The variance is nicely reduced to 0.8273 when $\kappa = 0.7$ and $c = 0.04$ is used on samples with the same separation and same number of observations.

Table 4.19 illustrates the estimation results for mixtures with medium separation ($r = 5$); whereas Table 4.20 presents the performance of the attenuated moment estimator in fitting mixtures where the components are well separated. We notice that the best combinations (κ, c) for small samples differ from the ones for large samples. We shall first analyse the results for large samples. Comparing these two tables, we are certain that the κ in the best combination decreases with an increase of the magnitude of the separation. For data sets with a large number of observations ($n_o = 1000$), the performance of this method is outstanding as the estimates are low in both bias and variances.

However, we found that the ideal combinations for small samples consist of a large fraction and a small attenuation, which are different from the ones for large samples. For example, for mixtures with $r = 10$ and $n_o = 10$, the best combination for estimating b is $\kappa = 1$ and $c = 0.09$; whereas when the number of observations is increased to 1000, the best combination becomes $\kappa = 0.2$ and $c = 0.04$. We investigated the 10000 estimates of b for the small samples and found that 28.55% of them are actually negative. These negative estimates have certainly reduced the variance of $\hat{b}$. Out of these 10000 estimates, 5.03% of $\hat{a}$ are negative; the proportion of negative $\hat{a}$ are relatively lower compared to $\hat{b}$. For samples of the same size, when $\kappa = 0.2$ and $c = 0.04$ is used, only 3.16% of $\hat{b}$ and 6.78% of $\hat{a}$ are

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	9.46×10^{-8} $(0.7, 0.03)$	4.49×10^{-9} $(0.2, 0.07)$	2.20×10^{-8} $(0.3, 0.01)$	1.37×10^{-8} $(0.8, 0.02)$	1.46×10^{-10} $(1, 0.06)$
$(\hat{b} - b)^2$	0.0055 $(1, 0.08)$	3.96×10^{-5} $(1, 0.05)$	3.46×10^{-5} $(1, 0.01)$	5.79×10^{-5} $(1, 0.03)$	2.13×10^{-6} $(0.5, 0.1)$
$(\hat{p} - p)^2$	3.07×10^{-9} $(0.6, 0.07)$	1.15×10^{-7} $(0.3, 0.02)$	5.64×10^{-6} $(0.2, 0.01)$	6.44×10^{-7} $(0.1, 0.1)$	9.83×10^{-10} $(0.7, 0.04)$
$Var[\hat{a}]$	0.0049 $(1, 0.01)$	0.0031 $(0.8, 0.01)$	0.0023 $(0.9, 0.01)$	0.0010 $(0.7, 0.01)$	4.16×10^{-5} $(0.3, 0.02)$
$Var[\hat{b}]$	291 $(1, 0.02)$	513 $(1, 0.04)$	489 $(0.8, 0.09)$	48 $(0.9, 0.08)$	0.0022 $(0.5, 0.05)$
$Var[\hat{p}]$	0.2302 $(0.5, 0.03)$	0.1446 $(1, 0.02)$	0.1197 $(0.9, 0.04)$	0.0551 $(0.9, 0.04)$	0.0030 $(0.7, 0.06)$
$MSE[\hat{a}]$	0.0050 $(0.9, 0.03)$	0.0031 $(0.8, 0.01)$	0.0023 $(0.9, 0.01)$	0.0010 $(0.7, 0.01)$	4.16×10^{-5} $(0.3, 0.02)$
$MSE[\hat{b}]$	291 $(1, 0.02)$	513 $(1, 0.04)$	491 $(0.8, 0.09)$	49 $(0.9, 0.08)$	0.0022 $(0.7, 0.09)$
$MSE[\hat{p}]$	0.2339 $(0.5, 0.03)$	0.1520 $(1, 0.02)$	0.1261 $(0.2, 0.01)$	0.0554 $(0.7, 0.03)$	0.0030 $(0.7, 0.06)$

Table 4.19: Performance of the method of attenuated rising factorial fractional moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	4.47×10^{-9} $(0.1, 0.08)$	2.45×10^{-7} $(0.1, 0.03)$	5.91×10^{-8} $(0.5, 0.07)$	3.42×10^{-11} $(0.8, 0.1)$	8.35×10^{-11} $(0.2, 0.03)$
$(\hat{b} - b)^2$	0.0108 $(1, 0.1)$	0.0142 $(1, 0.05)$	0.0027 $(1, 0.09)$	0.0009 $(1, 0.07)$	3.39×10^{-7} $(0.1, 0.1)$
$(\hat{p} - p)^2$	4.47×10^{-6} $(0.2, 0.03)$	3.06×10^{-8} $(0.1, 0.02)$	6.09×10^{-8} $(0.4, 0.07)$	1.40×10^{-8} $(0.5, 0.1)$	9.41×10^{-12} $(0.3, 0.04)$
$Var[\hat{a}]$	0.0055 $(0.4, 0.01)$	0.0032 $(0.3, 0.01)$	0.0021 $(0.4, 0.02)$	0.0006 $(0.2, 0.01)$	2.51×10^{-5} $(0.1, 0.01)$
$Var[\hat{b}]$	405 $(1, 0.09)$	635 $(0.9, 0.1)$	290 $(1, 0.09)$	0.1446 $(0.4, 0.08)$	0.0018 $(0.2, 0.04)$
$Var[\hat{p}]$	0.1020 $(0.5, 0.01)$	0.0673 $(0.7, 0.03)$	0.0498 $(0.6, 0.01)$	0.0209 $(0.4, 0.06)$	0.0009 $(0.2, 0.03)$
$MSE[\hat{a}]$	0.0057 $(0.9, 0.05)$	0.0033 $(0.9, 0.06)$	0.0022 $(0.4, 0.02)$	0.0006 $(0.2, 0.01)$	2.51×10^{-5} $(0.1, 0.01)$
$MSE[\hat{b}]$	405 $(1, 0.09)$	638 $(0.9, 0.1)$	290 $(1, 0.09)$	0.1486 $(0.4, 0.08)$	0.0018 $(0.3, 0.09)$
$MSE[\hat{p}]$	0.1081 $(0.5, 0.01)$	0.0677 $(0.7, 0.03)$	0.0501 $(0.6, 0.01)$	0.0210 $(0.4, 0.06)$	0.0009 $(0.2, 0.03)$

Table 4.20: Performance of the method of attenuated rising factorial fractional moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 0.6$ for different sample size n_o.

negative. However, with a lower fraction, b is more likely to be over-estimated if x_2 in (4.46) are under-estimated. Therefore, some over-estimated $\hat{b}$ caused $Var\left[\hat{b}\right]$ to become extremely large when a small fraction is used. Since a and b are probabilities, for both sets of estimates, we excluded unreasonable $\hat{a}$'s and $\hat{b}$'s (i.e. $\hat{a}$'s and $\hat{b}$'s which are outside the interval $(0,1)$) and drew the scatter plots of $\hat{b}$ against $\hat{a}$ in Figure 4.13. As a consequence, we are left with 4192 estimates given by $\kappa = 1$ and $c = 0.09$, and 5345 estimates given by $\kappa = 0.2$ and $c = 0.04$. It is now obvious from the scatter plots that the combination $(0.2, 0.04)$ provides a larger number of reasonable estimates compared to $(1, 0.09)$, the combination which appears to minimise the variance of $\hat{b}$ for small samples. In Figure 4.14, we draw the scatter plots of $\hat{b}$ against $\hat{a}$ when the number of observations is increased to 1000. Undoubtedly, $(0.2, 0.04)$ provides more reasonable estimates with lower variation, compared to the estimates given by $(1, 0.09)$.

Therefore, we should use the combination (κ, c) suggested by the large samples to estimate the parameters, although we might obtain an over-estimated $\hat{b}$ when the sample size is small if x_2 is badly estimated. However, the probability of getting a bad $\hat{x}_2$ will decrease when the number of observations is increased. Based on our simulation results, we should use a large fraction with a small attenuation for mixtures where the components have little separation. The magnitude of fraction should decrease when the degree of separation gets larger.

4.4.3 Asymptotic Covariance Matrix of the Attenuated Rising Factorial Fractional Moment Estimators

It will be meaningful if we know the variance of the attenuated rising factorial moment estimator for a mixture of two geometric distributions as it allows us to find the optimal combination of κ and c. We follow the procedure in Section 1.5.5 and find an approximation for the asymptotic covariance matrix $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ of the attenuated rising factorial fractional moment estimator $\boldsymbol{\Theta} = (a, b, p)$, which is given by

$$\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right] \approx \boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1} \boldsymbol{V}\left[\hat{\boldsymbol{\rho}}\right] \left(\boldsymbol{D}\left[\boldsymbol{\Theta}\right]^{-1}\right)^{T}. \tag{4.54}$$

Covariance Matrix of the Attenuated Rising Factorial Fractional Moments

We first find the covariance matrix of the attenuated rising factorial fractional moments $\boldsymbol{V}\left[\hat{\boldsymbol{\rho}}\right]$ with entries

$$Cov\left[\hat{\rho}_{\kappa}(c), \hat{\rho}_{\ell}(c)\right] = \frac{1}{n_o}\left[E\left[(N)_{\kappa}\varepsilon^{N}(N)_{\ell}\varepsilon^{N}\right] - E\left[(N)_{\kappa}\varepsilon^{N}\right]E\left[(N)_{\ell}\varepsilon^{N}\right]\right]. \tag{4.55}$$

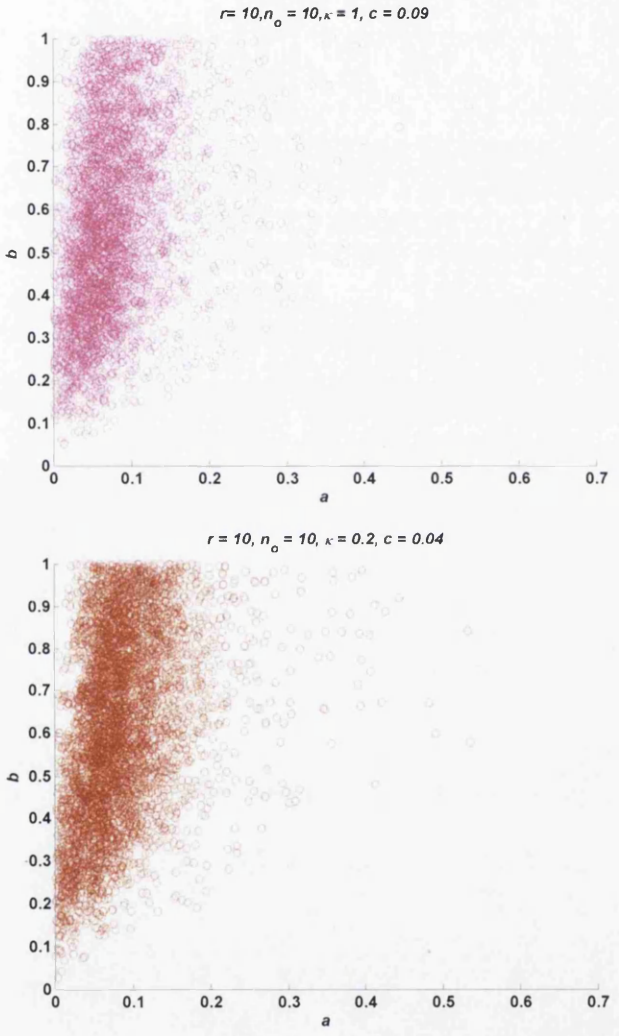


Figure 4.13: Scatter plots of $\hat{b}$ versus $\hat{a}$: Comparison of the performance of two different combinations of (κ, c) on geometric mixtures with small sample size.

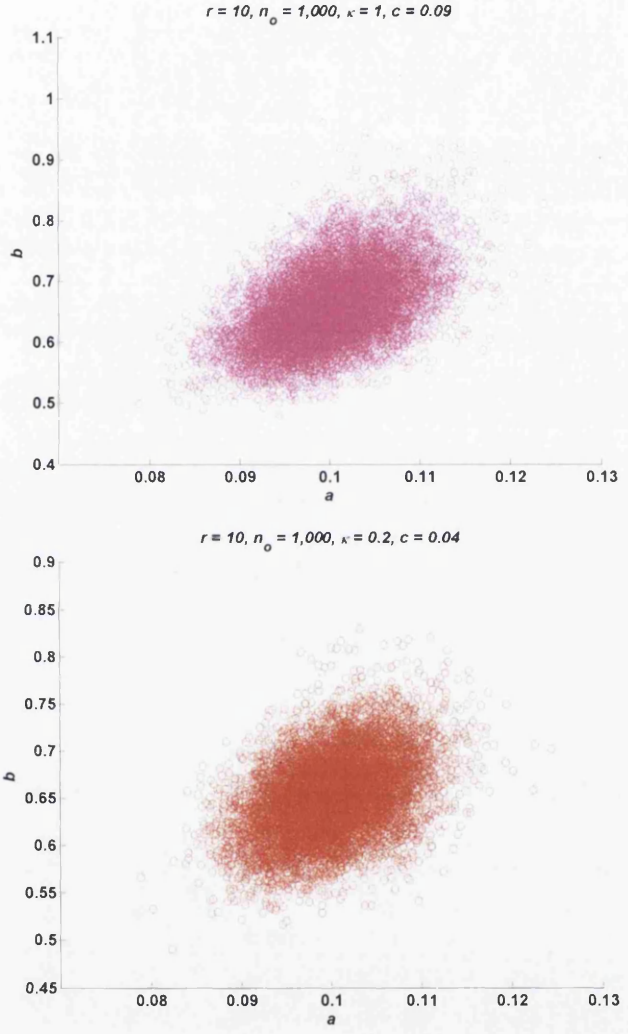


Figure 4.14: Scatter plots of $\hat{b}$ versus $\hat{a}$: Comparison of the performance two different combinations of (κ, c) on geometric mixtures with large sample size.

We denote $\xi(\kappa, \ell, c \mid a)$ as the expectation of the product of $(N)_\kappa \varepsilon^N$ and $(N)_\ell \varepsilon^N$ for a single geometric distribution with probability a. This expectation is defined as

$$\begin{aligned} \xi(\kappa, \ell, c \mid a) &= E\left[\frac{\Gamma(N+\kappa)\Gamma(N+\ell)}{\Gamma(N)^2}\varepsilon^{2N}\right] \\ &= a\varepsilon^2\Gamma(\kappa+1)\Gamma(\ell+1)\sum_{n=1}^{\infty}\left[(1-a)\varepsilon^2\right]^{n-1}\frac{(\kappa+1)_{n-1}(\ell+1)_{n-1}}{(1)_{n-1}(n-1)!}. \end{aligned} \tag{4.56}$$

The power series involved in (4.56) is again the Gaussian Hypergeometric function with parameters $(\kappa, \ell; 1)$. Hence, (4.56) is given by

$$\xi(\kappa, \ell, c \mid a) = a\varepsilon^2\Gamma(\kappa+1)\Gamma(\ell+1)_2F_1\left(\kappa+1, \ell+1; 1 \mid (1-a)\varepsilon^2\right). \tag{4.57}$$

Once again, we make use of the Kummer's solution and rewrite (4.57) in terms of a favourable convergent series:

$${}_2F_1\left(\kappa+1, \ell+1; 1 \mid (1-a)\varepsilon^2\right) = \left(1-(1-a)\varepsilon^2\right)^{-\kappa-\ell-1}{}_2F_1\left(-\kappa, -\ell; 1 \mid (1-a)\varepsilon^2\right). \tag{4.58}$$

Hence, we express $\xi(\kappa, \ell, c \mid a)$ using (4.58) as

$$\xi(\kappa, \ell, c \mid a) = a\varepsilon^2\Gamma(\kappa+1)\Gamma(\ell+1)\left(1-(1-a)\varepsilon^2\right)^{-\kappa-\ell-1}{}_2F_1\left(-\kappa, -\ell; 1 \mid (1-a)\varepsilon^2\right). \tag{4.59}$$

Using (4.59), we can now find the expectation of the product of $(N)_\kappa \varepsilon^N$ and $(N)_\ell \varepsilon^N$ for a mixture of two geometric distributions with parameter vector $\boldsymbol{\Theta} = (a, b, p)$ as follows:

$$\begin{aligned} E[(N)_\kappa \varepsilon^N (N)_\ell \varepsilon^N] &= pE\left[(N)_\kappa (N)_\ell \varepsilon^{2N} \mid a\right] + (1-p)E\left[(N)_\kappa (N)_\ell \varepsilon^{2N} \mid b\right] \\ &= \Gamma(\kappa+1)\Gamma(\ell+1)\varepsilon^2\left[\begin{array}{c} \dfrac{pa\ {}_2F_1\left(-\kappa, -\ell; 1 \mid (1-a)\varepsilon^2\right)}{\left(1-(1-a)\varepsilon^2\right)^{\kappa+\ell+1}} \\ +\dfrac{(1-p)b\ {}_2F_1\left(-\kappa, -\ell; 1 \mid (1-b)\varepsilon^2\right)}{\left(1-(1-b)\varepsilon^2\right)^{\kappa+\ell+1}} \end{array}\right]. \end{aligned} \tag{4.60}$$

Hence, substituting (4.60) and (4.39) into (4.55) we obtain the covariance between $\hat{\rho}_\kappa(c)$

and $\hat{\rho}_\ell(c)$ as

$$Cov\left[\hat{\rho}_\kappa(c),\hat{\rho}_\ell(c)\right]=\frac{1}{n_o}\varepsilon^2\Gamma(\kappa+1)\Gamma(\ell+1)\left[\begin{array}{c}\dfrac{pa_2F_1\left(-\kappa,-\ell;1\mid(1-a)\varepsilon^2\right)}{(1-(1-a)\varepsilon^2)^{\kappa+\ell+1}}\\ +\dfrac{qb_2F_1\left(-\kappa,-\ell;1\mid(1-b)\varepsilon^2\right)}{(1-(1-b)\varepsilon^2)^{\kappa+\ell+1}}\\ -\left(\begin{array}{c}\dfrac{p^2a^2}{(1-\varepsilon(1-a))^{\kappa+\ell+2}}\\ +\dfrac{q^2b^2}{(1-\varepsilon(1-b))^{\kappa+\ell+2}}\\ \dfrac{pqab}{(1-\varepsilon(1-a))^{\kappa+1}(1-\varepsilon(1-b))^{\ell+1}}\\ +\dfrac{pqab}{(1-\varepsilon(1-b))^{\kappa+1}(1-\varepsilon(1-a))^{\ell+1}}\end{array}\right)\end{array}\right]. \tag{4.61}$$

Therefore, the ij^{th} entry of the covariance matrix of the attenuated rising factorial fractional moments $\boldsymbol{V}[\hat{\boldsymbol{\rho}}]$ is given by $Cov\left[\hat{\rho}_i(c),\hat{\rho}_j(c)\right]$ according to the general expression in (4.61).

Jacobian Matrix of the Attenuated Rising Factorial Fractional Moment Estimator

Let $\boldsymbol{D}[\boldsymbol{\Theta}]$ be the Jacobian Matrix with entries: $d_{ij}=\dfrac{\partial\rho_{\kappa_i}(c)}{\partial\Theta_j}$, where

$$\frac{\partial\rho_{\kappa_i}(c)}{\partial a}=p\Gamma(\kappa_i+1)\left[\frac{\varepsilon}{(1-(1-a)\varepsilon)^{\kappa_i+1}}-\frac{\varepsilon^2a(\kappa_i+1)}{(1-(1-a)\varepsilon)^{\kappa_i+2}}\right], \tag{4.62}$$

$$\frac{\partial\rho_{\kappa_i}(c)}{\partial b}=q\Gamma(\kappa_i+1)\left[\frac{\varepsilon}{(1-(1-b)\varepsilon)^{\kappa_i+1}}-\frac{\varepsilon^2b(\kappa_i+1)}{(1-(1-b)\varepsilon)^{\kappa_i+2}}\right] \tag{4.63}$$

and

$$\frac{\partial\rho_{\kappa_i}(c)}{\partial p}=\varepsilon\Gamma(\kappa_i+1)\left[\frac{a}{(1-\varepsilon(1-a))^{\kappa_i+1}}-\frac{b}{(1-\varepsilon(1-b))^{\kappa_i+1}}\right] \tag{4.64}$$

for $i=1,2,3$. For simplicity, we set κ_2 as 2κ and κ_3 and 3κ for our investigation.

Approximated Asymptotic Covariance Matrix of the Attenuated Rising Factorial Fractional Moment Estimator

Now, we can find the approximated covariance matrix of the rising factorial fractional moment estimator by substituting the covariance matrix of the rising factorial fractional moments $\boldsymbol{V}[\hat{\boldsymbol{\rho}}]$ using (4.61) and the Jacobian matrix $\boldsymbol{D}[\boldsymbol{\Theta}]$ into (4.54). By substituting the true values of $\boldsymbol{\Theta}=(a,b,p)$ and n_o into (4.54), we obtain an approximation for the theoretical variance of this estimator.

With the method of attenuated rising factorial fractional moments, we use four moments (as shown in (4.41)) to estimate the three parameters of a mixture of two geometric distributions. Therefore, $\boldsymbol{D}[\boldsymbol{\Theta}]$ is a 4×3 matrix; in other words, we are not able to find the

inverse of $\boldsymbol{D}[\boldsymbol{\Theta}]$ as required in (4.54) because $\boldsymbol{D}[\boldsymbol{\Theta}]$ is not a square matrix. As discussed in the previous chapter, there are two available solutions to solve this problem.

Our previous work has shown that, although q does not need to be estimated, the agreement between the theoretical $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ and the practical variances of the estimators is sound when we add an extra column $\frac{\partial \mu_{\kappa_i}(c)}{\partial q}$ to the Jacobian matrix. We named this approach the Q-Version of the theoretical covariance matrix. Similarly, for the discrete analogue, we add

$$\frac{\partial \rho_{\kappa_i}(c)}{\partial q} = \varepsilon \Gamma(\kappa+1) \frac{b}{(1-\varepsilon(1-b))^{\kappa+1}} \tag{4.65}$$

to $\boldsymbol{D}[\boldsymbol{\Theta}]$ and amend $\frac{\partial \rho_{\kappa_i}(c)}{\partial p}$ to

$$\frac{\partial \rho_{\kappa_i}(c)}{\partial p} = \varepsilon \Gamma(\kappa+1) \frac{a}{(1-\varepsilon(1-a))^{\kappa+1}}. \tag{4.66}$$

Hence, $\boldsymbol{D}[\boldsymbol{\Theta}]$ is now a square matrix and invertible.

The second approach is the "GI-Version", in which we find the generalised inverse of $\boldsymbol{D}[\boldsymbol{\Theta}]$, given by

$$\boldsymbol{D}[\boldsymbol{\Theta}]^{-1} = \left(\boldsymbol{D}[\boldsymbol{\Theta}]^T \boldsymbol{D}[\boldsymbol{\Theta}]\right)^{-1} \boldsymbol{D}[\boldsymbol{\Theta}]^T \tag{4.67}$$

and

$$\left(\boldsymbol{D}[\boldsymbol{\Theta}]^T\right)^{-1} = \boldsymbol{D}[\boldsymbol{\Theta}] \left(\boldsymbol{D}[\boldsymbol{\Theta}]^T \boldsymbol{D}[\boldsymbol{\Theta}]\right)^{-1}. \tag{4.68}$$

Therefore, $\boldsymbol{D}[\boldsymbol{\Theta}]^{-1}$ is a 3×4 matrix and $\left(\boldsymbol{D}[\boldsymbol{\Theta}]^T\right)^{-1}$ becomes a 4×3 matrix; doing this makes $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$ a 3×3 matrix. Our previous investigation shows that the Q-Version outperforms the GI-Version by providing a stronger agreement to the practical results. Like the exponential case, the GI-Version is not a very reliable method for the geometric mixture to provide the theoretical variances with good accuracy. We have studied and found that the approximated theoretical variances of the attenuated rising factorial fractional estimators given by the GI-version for a mixture of two geometric distributions again is not as accurate as the Q-Version and hence we decided to only use the Q-Version for our study of the best combination of fraction and attenuation that gives the minimum variance of estimator.

4.4.4 Optimal κ and c

In this section, we find the optimal combination of fraction κ and attenuation c which gives the lowest variance of the estimator for a, b and p for a mixture of two geometric distributions. Tables 4.21, 4.22 and 4.23 allow us to check the conformity between the theoretical variance and the practical variance. In the second column of each table, we show the theoretical minimum variance of the estimator according to the Q-Version and their corresponding κ and c. In the last column, we first show the minimum practical variance

r	Theoretical $Var[\hat{a}]_Q$	Practical $Var[\hat{a}]$
2	**0.0003** **($\kappa = 0.95, c = 0.028$)**	0.0003 ($\kappa = 0.90, c = 0.010$) **0.0004** **($\kappa = 0.95, c = 0.028$)**
5	**4.11×10^{-5}** **($\kappa = 0.40, c = 0.022$)**	4.16×10^{-5} ($\kappa = 0.30, c = 0.020$) **4.14×10^{-5}** **($\kappa = 0.40, c = 0.022$)**
10	**2.51×10^{-5}** **($\kappa = 0.08, c = 0.008$)**	2.51×10^{-5} ($\kappa = 0.10, c = 0.010$) **2.51×10^{-5}** **($\kappa = 0.08, c = 0.008$)**

Table 4.21: Comparison of the theoretical and practical optimal combination of fraction and attenuation which gives the minimum variance of a using the method of attenuated rising factorial moments on a mixture of two geometric distributions with $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

found from our simulation experiment in Section 4.4.2 and their counterparts (κ, c) on the top part. For each optimal combination of κ and c suggested by the Q-Version, we simulated another 10000 data sets and estimated from each data set with the suggested combination, the variance obtained is recorded at the bottom part of the last column.

In Table 4.21, we can see that when $r = 2$, the theoretical optimal combination returns $Var[\hat{a}]$ which is marginally larger than the one obtained with $\kappa = 0.9$ and $c = 0.01$. Nevertheless, when r increases, the agreement is more sound and the theoretical optimal combination does give lower $Var[\hat{a}]$ than the ones obtained from our simulation experiments. It is clear from Tables 4.22 and 4.23 that, except for $r = 2$, the conformity between the theoretical and the practical variances of the estimators is excellent. Therefore, we can use the suggested theoretical combination of κ and c to estimate from a data set arising from a two-component geometric mixture distribution to enhance the precision of the estimates.

Let us now take a glance at the shape of $Var[\boldsymbol{\Theta}]_Q$ in Figures 4.15, 4.16 and 4.17. In each figure we see six graphs of $Var[\boldsymbol{\Theta}]_Q$ versus c; there are 5 lines in each graph representing a different fraction κ. We consider three different degrees of separation $r = (2, 5, 10)$ and for each r we consider 10 values of κ. From Figure 4.15, it is clear that we should use a large κ with a small c to obtain a highly precise estimate of a when the separation between the two component is small ($r = 2$). Both the best values of κ and c for estimating a decrease as the components become closer and closer to each other.

In Figure 4.16, we can see that when r increases, the best κ decreases but the ideal value of c increases (unlike its behaviour in the estimation of a). From Figure 4.17, it is clear that the optimal κ decreases for mixtures with a larger separation between the components whereas the best c has a value between 0.03 and 0.04.

r	Theoretical $Var\left[\hat{b}\right]_Q$	Practical $Var\left[\hat{b}\right]$
2	**0.0041** **($\kappa = 1.01, c = 0.03$)**	0.4134 ($\kappa = 0.70, c = 0.040$) **4** **($\kappa = 1.01, c = 0.033$)**
5	**0.0021** **($\kappa = 0.66, c = 0.061$)**	0.0022 ($\kappa = 0.50, c = 0.050$) **0.0022** **($\kappa = 0.66, c = 0.061$)**
10	**0.0018** **($\kappa = 0.21, c = 0.050$)**	0.0018 ($\kappa = 0.20, c = 0.040$) **0.0017** **($\kappa = 0.21, c = 0.050$)**

Table 4.22: Comparison of the theoretical and practical optimal combination of fraction and attenuation which gives the minimum variance of b using the method of attenuated rising factorial moments on a mixture of two geometric distributions with $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

r	Theoretical $Var\left[\hat{p}\right]_Q$	Practical $Var\left[\hat{p}\right]$
2	**0.1179** **($\kappa = 0.98, c = 0.031$)**	0.0711 ($\kappa = 0.80, c = 0.010$) **0.0719** **($\kappa = 0.98, c = 0.031$)**
5	**0.0029** **($\kappa = 0.51, c = 0.041$)**	0.0030 ($\kappa = 0.70, c = 0.060$) **0.0029** **($\kappa = 0.51, c = 0.041$)**
10	**0.0009** **($\kappa = 0.098, c = 0.030$)**	0.0009 ($\kappa = 0.20, c = 0.030$) **0.0009** **($\kappa = 0.098, c = 0.030$)**

Table 4.23: Comparison of the theoretical and practical optimal combination of fraction and attenuation which gives the minimum variance of p using the method of attenuated rising factorial moments on a mixture of two geometric distributions with $a = 0.1$, $b = 1 - 0.9^r$, $p = 0.6$ and $n_o = 1000$. Simulated figures are based on 10000 replications.

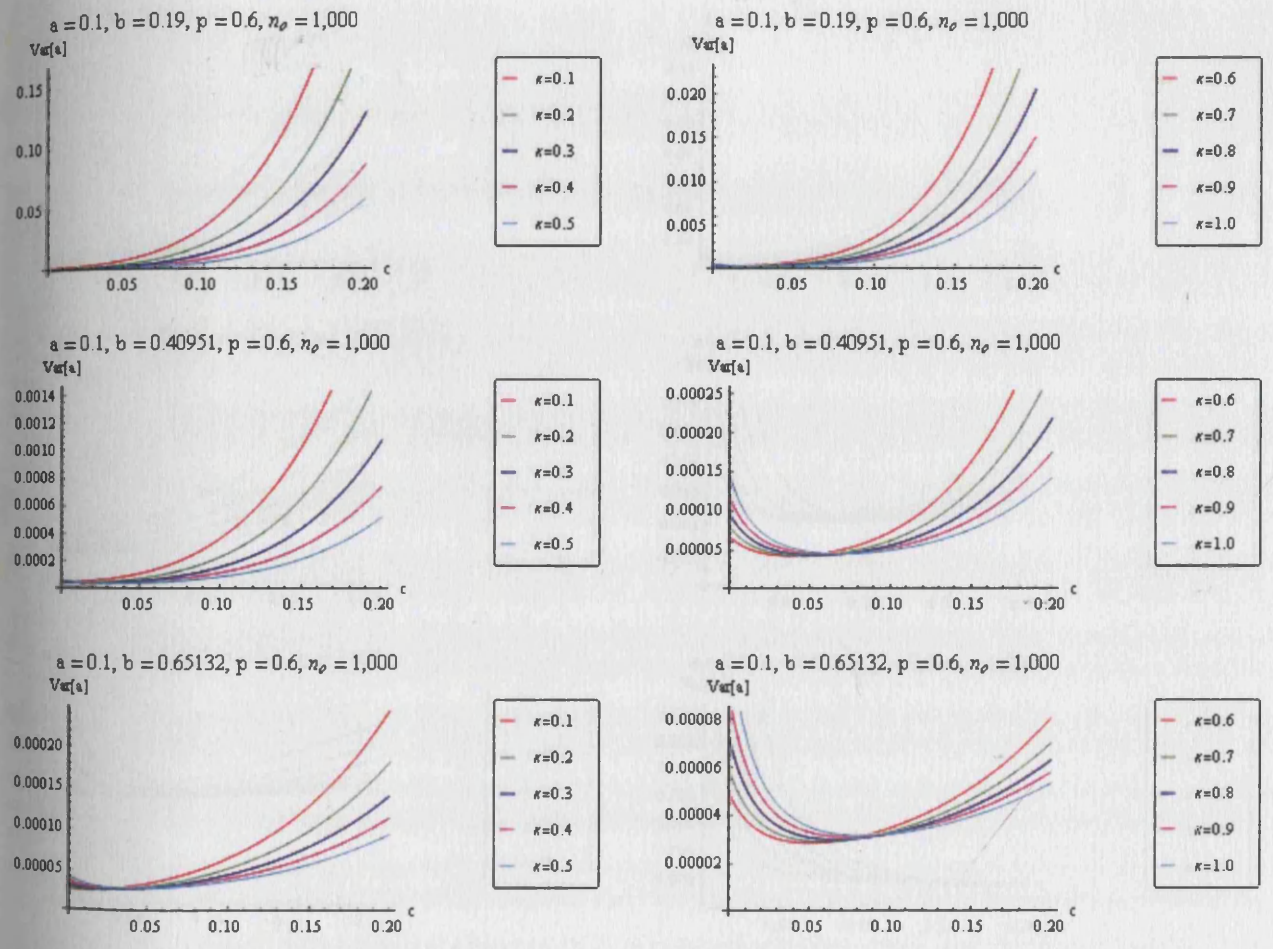


Figure 4.15: Plots of theoretical $Var\left[\hat{a}\right]_Q$ versus c for varying κ and r.

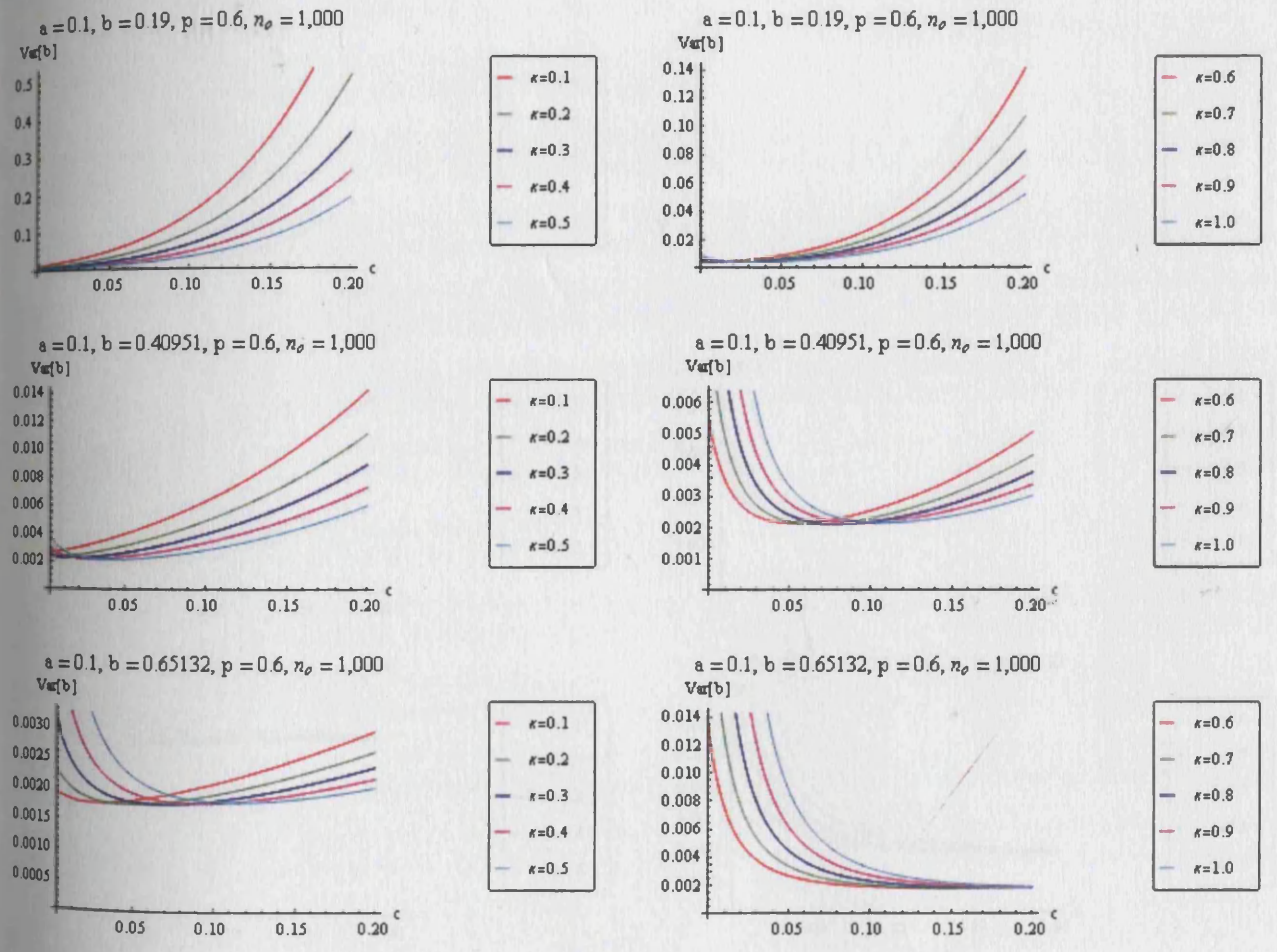


Figure 4.16: Plots of theoretical $Var\left[\hat{b}\right]_Q$ versus c for varying κ and r.

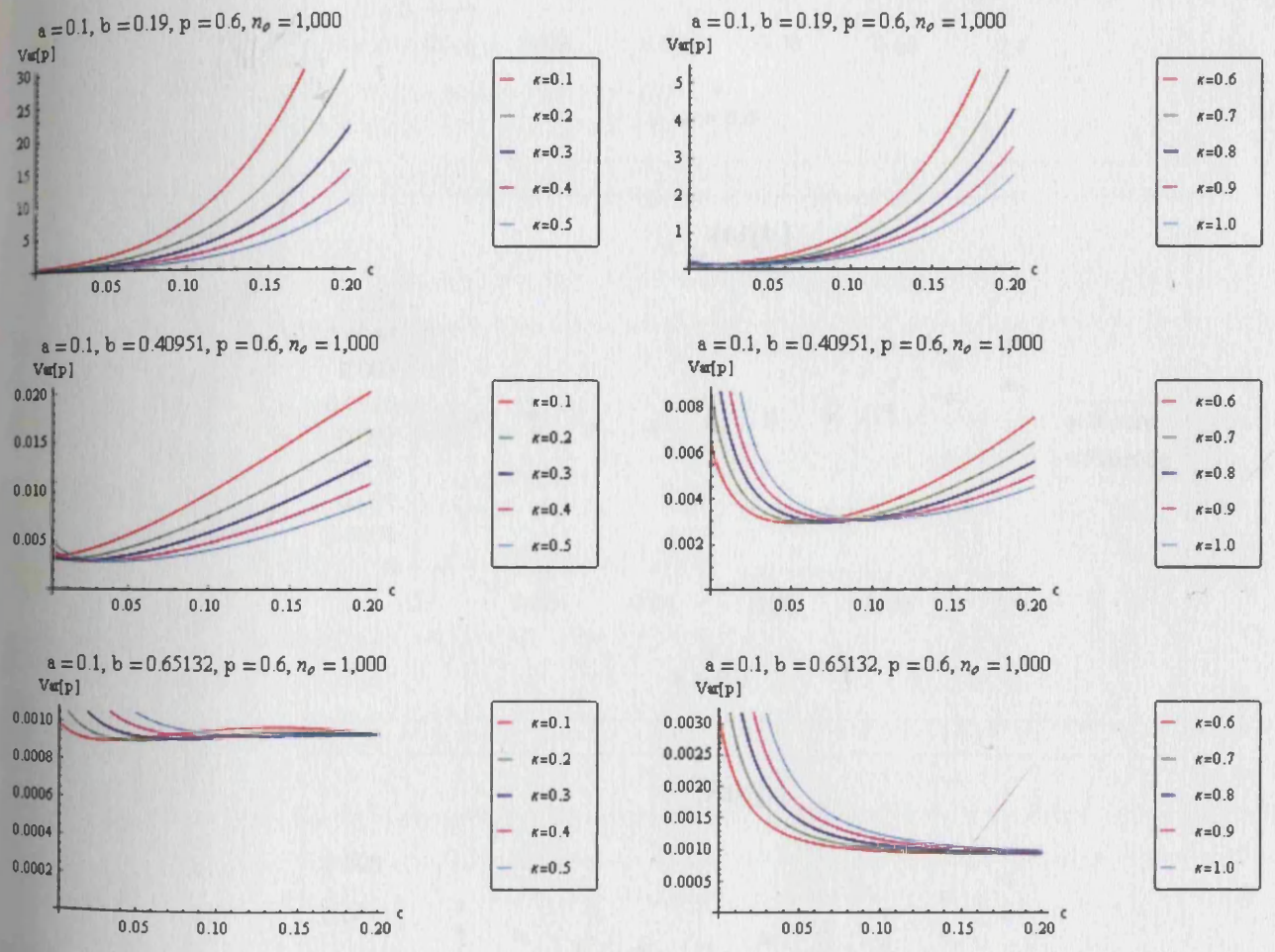


Figure 4.17: Plots of theoretical $Var\left[\hat{p}\right]_Q$ versus c for varying κ and r.

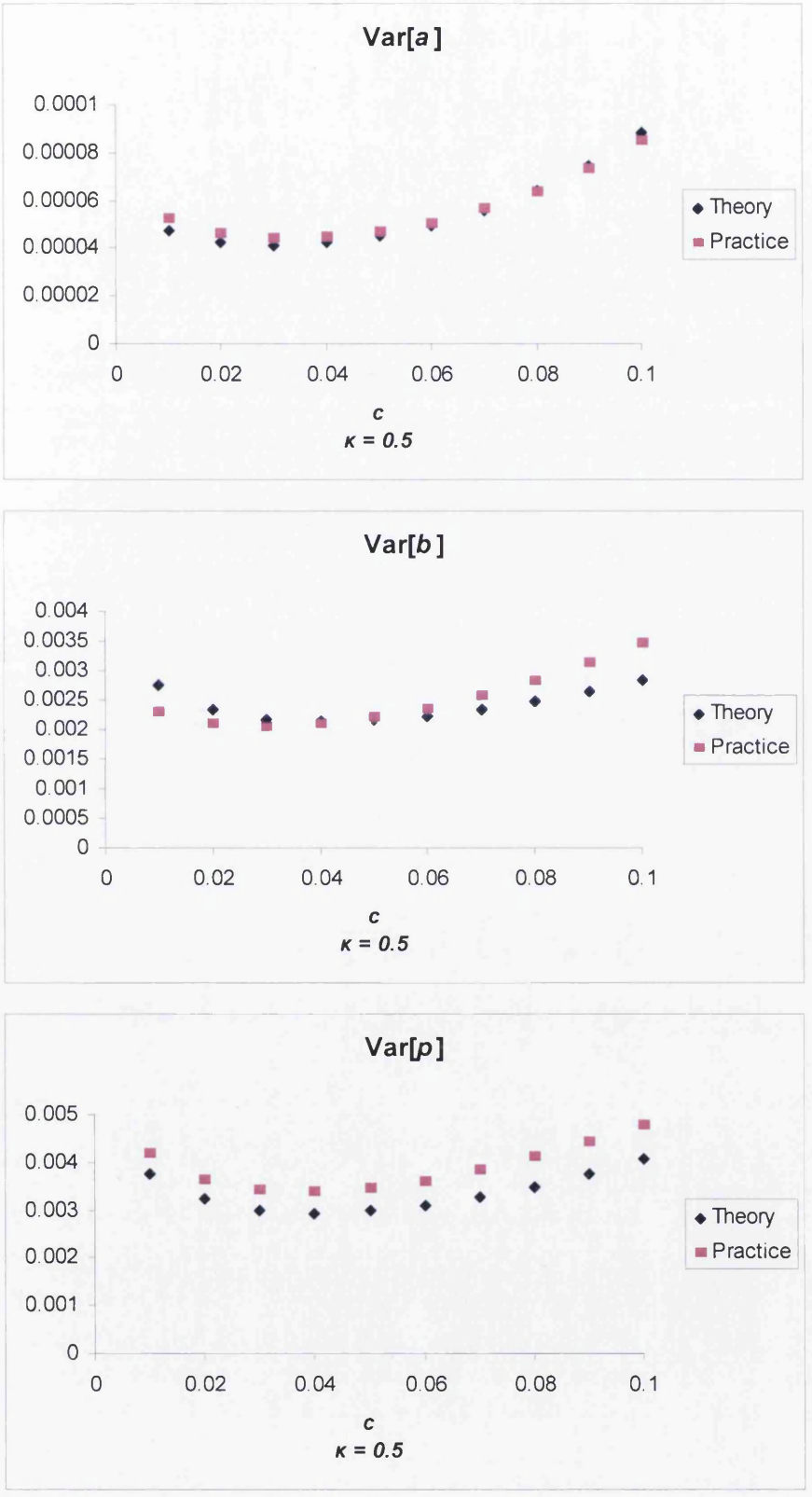


Figure 4.18: Asymptotic variance of the attenuated moment estimator given by true parameters and parameter estimates (estimated with $\kappa = 0.5$) versus c, based on a data set, consisting of 1000 observations, simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$.

By now, we know the best combinations of κ and c for estimating the parameters of geometric mixtures with different degrees of separation between the components. Unfortunately, in practice, we do not know the degree of separation. We shall now suggest a way for users to choose the best estimates given by the attenuated moment estimator. We simulated a data set, consisting of 1000 observations, from a mixture of two geometric distributions with true parameter $a = 0.1$, $b = 0.4095$ and $p = 0.6$, and estimated the parameters with ten combinations of κ and c, where κ is fixed at 0.5 and c varies from 0.01 to 0.1. We then substituted the yielded sets of parameter estimates into (4.54) and plotted the variances alongside the theoretical variances given by the true parameters versus c in Figure 4.18. Since the conformity between the theory and practice is satisfactory, we suggest users estimate the parameters with a number of combinations of κ and c in practice. The best set of estimates should have the minimum $Var\left[\hat{\boldsymbol{\Theta}}\right]$ when they are substituted into (4.54).

4.4.5 Discussion

This section has demonstrated, in both theory and simulation, that the method of attenuated rising factorial fractional moments is a robust method for estimating the parameters in a mixture of two geometric distributions. It provides parameter estimates with significantly lower bias and variances than the ones given by the standard moment estimator and the fractional moment estimator. As expected, results are more promising when components are better separated or the sample size is sufficiently large. Of course, like any modified moment estimator, this method does not guarantee the estimates to lie in the interval $(0, 1)$. Although the estimation of b remains problematic for samples of very small sizes, we did find good estimates of b, which are lowly biased and have small variances, for samples with only fifty observations when the components in a mixture are well separated, as shown in Table 4.20. We also obtained the approximated asymptotic variances of the estimators which have a strong conformity with the simulated values. Therefore, we are able to suggest ideal combinations of κ and c for estimating the parameters in geometric mixtures with different degrees of separation. In real life, we do not know the separation between the components. However, if one estimate the parameters with a few different combinations of κ and c, the best set of estimates should be the ones which minimises $Var\left[\hat{\boldsymbol{\Theta}}\right]$ when they are substituted into (4.54).

4.5 The Method Based on an Appell Sequences

4.5.1 Introduction

We have seen how an Appell sequences can be used to estimate the parameters of a mixture of two exponential distributions. In this section, we study a new method constructed by Jalali (2005a) which is based on an Appell double sequence for the parameter estimation of a two-component geometric mixture distribution. Our exposition in this subsection is based

on his paper. The geometric random variables are assumed to have non-zero probabilities on the whole of natural numbers including 0.

Let $\{a_k(n) : k = 0, 1, ...; n = 0, \pm 1, ...\}$ be a double sequence such that for all k,

$$a_k(n) = 0 \tag{4.69}$$

when n is negative and

$$\sum_{n=0}^{\infty} a_k(n)\,\bar{\theta}^n \tag{4.70}$$

exists for all $\bar{\theta}$ in the interior of the unit disc in the complex plane.

We let

$$A_k(\bar{\theta}) = (1-\bar{\theta}) \sum_{n=0}^{\infty} a_k(n)\,\bar{\theta}^n, \tag{4.71}$$

and define the operator φ acting on sequences in the following way:

$$\varphi c(n) = c(n+1). \tag{4.72}$$

Then (4.71) becomes

$$A_k(\bar{\theta}) = (1-\bar{\theta}) \sum_{n=0}^{\infty} \bar{\theta}^n \varphi^n a_k(0). \tag{4.73}$$

Note that

$$\varphi^{-1} c(n) = c(n-1) \tag{4.74}$$

by convention. We then define the difference operators $\Delta = \varphi - 1$ and $\nabla = 1 - \varphi^{-1}$ and set

$$b_k(n) = \nabla a_k(n). \tag{4.75}$$

With this definition we have

$$\begin{aligned} B_k(\bar{\theta}) &= (1-\bar{\theta}) \sum_{n=0}^{\infty} \bar{\theta}^n \varphi^n b_k(0) \qquad (4.76)\\ &= (1-\bar{\theta}) \sum_{n=0}^{\infty} \bar{\theta}^n \left(\varphi^n - \varphi^{n-1}\right) a_k(0) \\ &= (1-\bar{\theta}) \sum_{n=0}^{\infty} \left(\bar{\theta}^n - \bar{\theta}^{n+1}\right) \varphi^n a_k(0) \\ &= (1-\bar{\theta})^2 \sum_{n=0}^{\infty} \bar{\theta}^n \varphi^n a_k(0). \end{aligned}$$

Substituting (4.73) into (4.76) yields

$$B_k(\bar{\theta}) = (1-\bar{\theta})\, A_k(\bar{\theta}). \tag{4.77}$$

Now we choose the double sequence $a_k(n)$ such as $a_{k-1}(n) = \nabla a_k(n)$, for all $k = 1, 2, ...$, then we have

$$A_k(\bar{\theta}) = (1-\bar{\theta})^{-1} A_{k-1}(\bar{\theta}),$$

and by induction,

$$A_k\left(\bar{\theta}\right) = \left(1-\bar{\theta}\right)^{-k} A_0\left(\bar{\theta}\right). \tag{4.78}$$

If we start by any sequence $\{a_0(n)\}$ satisfying the aforementioned conditions, then

$$a_1(n) = \nabla^{-1} a_0(n),$$

and hence

$$\begin{aligned} a_1(n) &= \left(1-\varphi^{-1}\right)^{-1} a_0(n) \\ &= \sum_{l=0}^{\infty} a_0(n-l) = \sum_{l=0}^{n} a_0(l). \end{aligned}$$

We note that

$$\begin{aligned} a_k(n) &= \left(1-\varphi^{-1}\right)^{-k} a_0(n) \\ &= \sum_{l=0}^{\infty} \binom{k+l-1}{l} \varphi^{-l} a_0(n) \\ &= \sum_{l=0}^{n} \binom{k+l-1}{l} a_0(n-l). \end{aligned} \tag{4.79}$$

Therefore, by choosing the initial sequence $\{a_0(n)\}$, we can construct the whole of our double sequence

$$\begin{aligned} a_k(n) &= \sum_{l=0}^{n} (-1)^l \binom{-k}{l} a_0(n-l) \\ &\Leftrightarrow a_k(n) = \sum_{l=0}^{n} \binom{k+l-1}{l} a_0(n-l) \\ &\Leftrightarrow a_k(n) = \sum_{l=0}^{n} \binom{k+n-l-1}{k-1} a_0(l). \end{aligned} \tag{4.80}$$

Clearly, these satisfy the recurrence relation $a_{k-1}(n) = \nabla a_k(n)$, and thus $A_k\left(\bar{\theta}\right) = \left(1-\bar{\theta}\right)^{-k} A_0\left(\bar{\theta}\right)$. We call such sequences Appell double sequences or Appell (infinite) matrices.

Practical Estimation

Suppose we have a mixture of two geometric distributions with

$$f(n;\boldsymbol{\Theta}) = pa(1-a)^n + (1-p)\,b(1-b)^n, \quad n = 0,1,2,...,n,...$$

where $\boldsymbol{\Theta} = (a,b,p)$. The Appell moments δ_k is given by

$$\delta_k = E\left[a_k(n)\right] = pA_k(1-a) + (1-p)\,A_k(1-b). \tag{4.81}$$

These identities are similar to the ones we obtained in the exponential case (see Section 3.5). Next suppose $\{n_i : i = 1, ..., n_o\}$ is a sample of size n_o from our mixture distribution, then the sample-expectations are as follows:

$$\hat{\delta}_k = \frac{1}{n_o} \sum_{i=1}^{n_o} a_k (n_i) .$$

Upon equating these with δ_k's we find the estimates of our parameters.

Examples of Appell Double Sequences

Example 1 *Let*

$$\begin{aligned} a_0 (n) &= 1 \text{ if } n \geq 0 \\ &= 0 \text{ otherwise,} \end{aligned}$$

then

$$A_0 (\bar{\theta}) = 1,$$

and thus our Appell moments are

$$\delta_k = pa^{-k} + (1-p) b^{-k}.$$

Clearly,

$$\begin{aligned} a_1 (n) &= n + 1, \\ a_k (n) &= \frac{(n+1)(n+2)...(n+k)}{k!} = \begin{pmatrix} n+k \\ k \end{pmatrix}, \end{aligned}$$

so our double sequences are rising factorials, this is the case of ordinary moments.

Example 2 *This is the above case in reverse. We let*

$$\begin{aligned} a_k (n) &= 1 \text{ if } n \geq 0 \\ &= 0 \text{ otherwise,} \end{aligned}$$

and find

$$\begin{aligned} a_{k-l}(n) &= \nabla^l a_k(n) \\ &= \left(1-\varphi^{-1}\right)^l a_k(n) \\ &= \sum_{i=0}^{l}(-1)^i \binom{l}{i} a_k(n-i) \\ &= \sum_{i=0}^{\min\{l,n\}}(-1)^i \binom{l}{i}. \end{aligned}$$

Clearly, when $n \geq l > 0$, *this is equal to zero. Hence,*

$$\begin{aligned} a_{k-l}(n) &= \sum_{i=0}^{n}(-1)^i \binom{l}{i} = (-1)^n \binom{l-1}{n} \quad \text{if } n < l \\ &= 0 \qquad\qquad \text{otherwise.} \end{aligned}$$

In particular,

$$\begin{aligned} a_{k-1}(0) &= 1,\ a_{k-1}(1) = a_{k-1}(2) = ... = 0; \\ a_{k-2}(0) &= 1,\ a_{k-2}(1) = -1,\ a_{k-2}(2) = a_{k-2}(3) = ... = 0; \\ a_{k-3}(0) &= 1,\ a_{k-3}(1) = -2,\ a_{k-3}(2) = 1,\ a_{k-3}(3) = a_{k-3}(4) = ... = 0. \end{aligned}$$

Note that in this case

$$\delta_{k-l} = pa^l + (1-p)\,b^l.$$

Example 3 *One can easily combine Example* 1 *and Example* 2. *For example when we have a mixture of three geometric distributions, we need at least five moments. We can start by letting*

$$\begin{aligned} a_0(0) &= 1,\ a_0(1) = -2,\ a_0(2) = 1,\ a_0(3) = a_0(4) = ... = 0; \\ a_1(0) &= 1,\ a_1(1) = -1,\ a_1(2) = a_1(3) = ... = 0; \\ a_2(0) &= a_2(1) = ... = 1; \\ a_3(n) &= n+1; \\ a_4(n) &= \frac{(n+1)(n+2)}{2}. \end{aligned}$$

In this particular case

$$\delta_l = \sum_{i=1}^{3} pa^{3-l} + (1-p)\,b^{3-1},\ l = 0,1,2,3,4.$$

Example 4 *Define the l^{th} Kronecker sequence as*

$$\begin{aligned} \varepsilon^{(l)}(n) &= 1 \text{ if } n = l \\ &= 0 \text{ otherwise} \end{aligned}$$

Clearly from Example 1, *we have*

$$\nabla^{-k}\varepsilon^{(l)}(n) = \begin{pmatrix} n-l+k-1 \\ n-l \end{pmatrix} = \begin{pmatrix} n-l+k-1 \\ k-1 \end{pmatrix}.$$

With this definition, we can now choose sequences which are dependent on our sample $\{n_i : i = 1, ..., n_o\}$. *As an example, we consider the empirical frequency function of our sample. This is the sequence* $a_0(n)$, *where the latter is the number of occurences of* n *in our sample. Then,*

$$a_0(n) = \sum_{i=1}^{n_o} \varepsilon^{(n_i)}(n) \tag{4.82}$$

and

$$a_k(n) = \nabla^{-k} a_0(n) = \sum_{i=1}^{n_o} \begin{pmatrix} n-n_i+k-1 \\ k-1 \end{pmatrix}. \tag{4.83}$$

Obviously, $a_1(n)$ *is the number of sample points not exceeding* n. *This, of course, is* n_o *times the ECDF of our sample.*

In an simulation experiment, we investigate the performance of the Appell moment estimator in estimating the parameters of a mixture of two geometric distributions, particularly with Krocnecker sequences (Example 4). We shall first show how such sequences can be used in practice to solve the estimation problem. In theory, given (4.82),

$$A_0(\bar{\theta}) = (1-\bar{\theta}) \sum_{i=1}^{n_o} \bar{\theta}^{n_i},$$

and following (4.78), we know

$$A_k(\bar{\theta}) = (1-\bar{\theta})^{-k+1} \sum_{i=1}^{n_o} \bar{\theta}^{n_i}.$$

Therefore, for a mixture of two geometric distributions with $\Theta = (a, b, p)$, from (4.78), the k^{th} theoretical moment is given by

$$\delta_k = pa^{-k+1} \sum_{i=1}^{n_o} (1-a)^{n_i} + (1-p)\, b^{-k+1} \sum_{i=1}^{n_o} (1-b)^{n_i}. \tag{4.84}$$

In particular, we use four moments to estimate the parameters. Their theoretical forms are as follows

$$\delta_0 = pa\sum_{i=1}^{n_o}(1-a)^{n_i} + (1-p)\,b\sum_{i=1}^{n_o}(1-b)^{n_i}\,, \tag{4.85}$$

$$\delta_1 = p\sum_{i=1}^{n_o}(1-a)^{n_i} + (1-p)\sum_{i=1}^{n_o}(1-b)^{n_i}\,, \tag{4.86}$$

$$\delta_2 = \frac{p}{a}\sum_{i=1}^{n_o}(1-a)^{n_i} + \frac{(1-p)}{b}\sum_{i=1}^{n_o}(1-b)^{n_i}\,, \tag{4.87}$$

$$\delta_3 = \frac{p}{a^2}\sum_{i=1}^{n_o}(1-a)^{n_i} + \frac{(1-p)}{b^2}\sum_{i=1}^{n_o}(1-b)^{n_i}\,. \tag{4.88}$$

If we let

$$I_0(\theta) = \theta\sum_{i=1}^{n_o}(1-\theta)^{n_i}\,,$$

$$w_1 = pI_0(a)\,,$$

and

$$w_2 = (1-p)\,I_0(b)\,,$$

then (4.84) can now be expressed as

$$\delta_k = w_1\mu_1^k + w_2\mu_2^k, \tag{4.89}$$

where $\mu_1 = a^{-1}$ and $\mu_2 = b^{-1}$. The estimates of μ_1 and μ_2 are the roots of the following quadratic equation

$$\det\begin{bmatrix} \hat{\delta}_0 & \hat{\delta}_1 & \hat{\delta}_2 \\ \hat{\delta}_1 & \hat{\delta}_2 & \hat{\delta}_3 \\ 1 & u & u^2 \end{bmatrix} = 0, \tag{4.90}$$

where $\hat{\delta}_k$'s are the raw sample moments. Hence, the estimates of a and b are given by

$$\begin{aligned} \hat{a} &= \frac{1}{\hat{\mu}_1}, \\ \hat{b} &= \frac{1}{\hat{\mu}_2}. \end{aligned} \tag{4.91}$$

To estimate p, in order to fully utilise all four moments, we define the following extended Vandermonde matrix

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 \\ \hat{\mu}_1 & \hat{\mu}_2 \\ \hat{\mu}_1^2 & \hat{\mu}_2^2 \\ \hat{\mu}_1^3 & \hat{\mu}_2^3 \end{bmatrix}, \tag{4.92}$$

and let

$$\mathbf{\Lambda} = diag\left[\begin{array}{cccc} \lambda_0 & \lambda_1 & \lambda_2 & \lambda_3 \end{array}\right]$$

be a diagonal matrix with positive weights on its main diagonal. Since (4.92) has no ordinary inverse and from (4.89) we know

$$\boldsymbol{V}\left[\begin{array}{c} w_1 \\ w_2 \end{array}\right] = \left[\begin{array}{c} \delta_0 \\ \delta_1 \\ \delta_2 \\ \delta_3 \end{array}\right],$$

we seek, then, to find values $\hat{w}_1$ and $\hat{w}_2$ such that the difference

$$\boldsymbol{V}\left[\begin{array}{c} \hat{w}_1 \\ \hat{w}_2 \end{array}\right] - \left[\begin{array}{c} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{array}\right]$$

be minimum in the sense of least squared sum with weights λ_0, ..., λ_1, which means that we shall find $\hat{w}_1$ and $\hat{w}_2$ which minimise the following weighted sum of squared errors

$$\left(\boldsymbol{V}\left[\begin{array}{c} \hat{w}_1 \\ \hat{w}_2 \end{array}\right] - \left[\begin{array}{c} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{array}\right]\right)^T \mathbf{\Lambda} \left(\boldsymbol{V}\left[\begin{array}{c} \hat{w}_1 \\ \hat{w}_2 \end{array}\right] - \left[\begin{array}{c} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{array}\right]\right). \tag{4.93}$$

As we have shown before, the vector of weights which makes (4.93) is

$$\left[\begin{array}{c} \hat{w}_1 \\ \hat{w}_2 \end{array}\right] = \left(\boldsymbol{V}^T \mathbf{\Lambda} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^T \mathbf{\Lambda} \left[\begin{array}{c} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{array}\right]. \tag{4.94}$$

Having found $\hat{w}_1$, $\hat{w}_2$, $\hat{a}$ and $\hat{b}$, the estimate of p can now be found by

$$p = \frac{\dfrac{\hat{w}_1}{I_0(\hat{a})}}{\dfrac{\hat{w}_1}{I_0(\hat{a})} + \dfrac{\hat{w}_2}{I_0\left(\hat{b}\right)}}. \tag{4.95}$$

We shall now demonstrate how the practical moments can be obtained given a raw sample, from (4.83), (4.84) can be estimated from

$$\begin{aligned}\hat{\delta}_k &= \frac{1}{n_o}\sum_{i=1}^{n_o} a_k(n_i) \\ &= \frac{1}{n_o}\sum_{i=1}^{n_o}\sum_{j=1}^{n_o}\binom{n_i - n_j + k - 1}{k-1}.\end{aligned}$$

In particular, $\hat{\delta}_0$ is given by

$$\hat{\delta}_0 = \frac{1}{n_o}\sum_{i=1}^{n_o}\sum_{j=1}^{n_o}\text{number of occurrences of } n_i = n_j, \tag{4.96}$$

$$\hat{\delta}_1 = \frac{1}{n_o}\sum_{i=1}^{n_o}\sum_{j=1}^{n_o}\text{number of occurrences of } n_i \geq n_j, \tag{4.97}$$

$$\hat{\delta}_2 = \frac{1}{n_o}\sum_{i=1}^{n_o}\sum_{n_i \geq n_j} n_i - n_j + 1, \tag{4.98}$$

and

$$\hat{\delta}_3 = \frac{1}{n_o}\sum_{i=1}^{n_o}\sum_{n_i \geq n_j}\frac{(n_i - n_j + 1)(n_i - n_j + 2)}{2}. \tag{4.99}$$

In the following example, we show how (4.96) to (4.99) can be obtained from a raw sample.

Example 5 *Assume we have a raw sample with three observations:* $n = \{1, 1, 2\}$. *To find the raw moments, this is what we do:*

1. *Find* $\hat{\delta}_0$.

   With the data, we form a matrix as follows

$$\begin{array}{c|ccc} \frac{n_j}{n_i} & 1 & 1 & 2 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{array},$$

   where the element e_{ij} *is* 1 *if* $n_i = n_j$. *Therefore,* $\hat{\delta}_0 = \frac{5}{3}$.

2. *Find* $\hat{\delta}_1$.

We form another matrix

$$\begin{array}{c} \frac{n_j}{n_i} \\ 1 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 1 & 2 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \end{array},$$

where the element e_{ij} *is* 1 *if* $n_i \geq n_j$. *Hence* $\hat{\delta}_1 = \frac{7}{3}$.

3. *Find* $\hat{\delta}_2$.

   The following matrix is formed

$$\begin{array}{c} \frac{n_j}{n_i} \\ 1 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 1 & 2 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix} \end{array},$$

   where the element e_{ij} *is* $n_i - n_j + 1$ *if* $n_i \geq n_j$. *Hence* $\hat{\delta}_2 = \frac{9}{3} = 3$.

4. *Finally, we find* $\hat{\delta}_3$.

   The matrix for $\hat{\delta}_3$ *is*

$$\begin{array}{c} \frac{n_j}{n_i} \\ 1 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 1 & 2 \end{array} \\ \frac{1}{2}\begin{pmatrix} 1+2 & 1+2 & 0 \\ 1+2 & 1+2 & 0 \\ 2+3 & 2+3 & 1+3 \end{pmatrix} \end{array},$$

   where the element e_{ij} *is* $\frac{(n_i - n_j + 1)(n_i - n_j + 2)}{2}$ *if* $n_i \geq n_j$. *Hence* $\hat{\delta}_2 = \frac{13}{3}$.

Upon substituting (4.96) to (4.99) into (4.90) and following procedures from (4.91) to (4.95), we then obtain estimates of a, b and p.

4.5.2 Simulation Results

Tables 4.24 to 4.26 show the performance of the method using Kronecker sequences for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 1 - (1-a)^r$ and $p = 0.6$ for three different ratios r and five different sample sizes n_o. We observe poor estimates of b and p for samples of small sizes and small separation between the two components (see Table 4.24 for $r = 2$). Nevertheless, this method is able to provide reasonable parameter estimates when the sample size is large enough ($n_o = 1000$); all parameter estimates are lowly biased and have small variances, except for $\hat{b}$ when the components are hardly distinguishable ($r = 2$). For large samples, the variances of $\hat{a}$ and $\hat{p}$

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\hat{a} - a\right)^2$	0.0026	0.0013	0.0008	0.0001	3.18×10^{-7}
$\left(\hat{b} - b\right)^2$	27	1.0077	0.0317	0.13463	0.0036
$\left(\hat{p} - p\right)^2$	0.0432	0.0224	0.0381	0.0215	0.0008
$Var\left[\hat{a}\right]$	0.0039	0.0024	0.0018	0.0010	0.0002
$Var\left[\hat{b}\right]$	120756	8999	1369	3892	10
$Var\left[\hat{p}\right]$	10	9	9	3	0.0604
$MSE\left[\hat{a}\right]$	0.0065	0.0037	0.0026	0.0011	0.0002
$MSE\left[\hat{b}\right]$	120795	9000	1369	3892	10
$MSE\left[\hat{p}\right]$	10	9	9	3	0.0612

Table 4.24: Performance of the method of Appell moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 0.6$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\hat{a} - a\right)^2$	0.0030	0.0014	0.0008	0.0002	5.70×10^{-7}
$\left(\hat{b} - b\right)^2$	0.0205	0.8913	0.0065	0.1051	0.0008
$\left(\hat{p} - p\right)^2$	0.0275	0.0246	0.0167	0.0074	4.29×10^{-5}
$Var\left[\hat{a}\right]$	0.0057	0.0034	0.00236	0.0010	6.55×10^{-5}
$Var\left[\hat{b}\right]$	2510	3220	2254	932	0.0069
$Var\left[\hat{p}\right]$	5	0.8824	0.2151	0.3324	0.0060
$MSE\left[\hat{a}\right]$	0.0087	0.0049	0.0032	0.00116	6.60×10^{-5}
$MSE\left[\hat{b}\right]$	2510	3221	2254	932	0.0076
$MSE\left[\hat{p}\right]$	5	0.9070	0.2318	0.3397	0.0060

Table 4.25: Performance of the method of Appell moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\bar{\hat{a}} - a\right)^2$	0.0029	0.0013	0.0008	0.0001	3.78×10^{-7}
$\left(\bar{\hat{b}} - b\right)^2$	1.1930	0.9093	0.9484	0.3345	0.0006
$\left(\bar{\hat{p}} - p\right)^2$	0.0228	0.0169	0.0145	0.0338	1.90×10^{-5}
$Var\left[\hat{a}\right]$	0.0063	0.0036	0.0025	0.0009	4.84×10^{-5}
$Var\left[\hat{b}\right]$	6149	8213	5573	3243	0.0095
$Var\left[\hat{p}\right]$	6	0.4592	0.1377	0.0490	0.0031
$MSE\left[\hat{a}\right]$	0.0092	0.0050	0.0032	0.0010	4.87×10^{-5}
$MSE\left[\hat{b}\right]$	6150	8214	5574	3243	0.0101
$MSE\left[\hat{p}\right]$	7	0.4762	0.1522	0.0550	0.0031

Table 4.26: Performance of the method of Appell moments for 10000 data sets simulated from a mixture of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 0.6$ for different sample size n_o.

decrease with an increase of the magnitude of the separation. However, the variance of $\hat{b}$ is lowest when $r = 5$.

4.5.3 Discussion

This is a new method which makes use of double Appell sequences to fit a discrete geometric mixture. It is exciting to observe reasonable parameter estimates provided by this method when a sample is large enough. Nevertheless, our simulation results show that its performance is not as good as the other methods studied in previous sections, especially when the number of observations in a sample is limited. We investigated the reasons for the poor estimation and found that $Var\left[\hat{\delta}_2\right]$ and $Var\left[\hat{\delta}_3\right]$ are, like the ordinary method of rising factorial moments, extremely high. For instance, when $r = 2$, $p = 0.6$ and $n_o = 1000$, $Var\left[\hat{\delta}_2\right]$ is 25879 and $Var\left[\hat{\delta}_3\right]$ is 13674315. Obviously, these large variances of moments have negative impacts on the precision of the parameter estimates.

We are free to use different Appell double sequences with this method and we demonstrated four examples here. Our investigation showed that the Kronecker sequence performs better than the other three Appell sequences. In the future, we can try other Appell double sequences to investigate whether or not the variances of the moments can be controlled and hence better estimates of b and p can be obtained.

4.6 Comparison of Estimation Methods

In this chapter, we have evaluated five different methods, the MLE via the EM algorithm, the method of rising factorial moments (MM), the method of rising factorial fractional moments (FM), the method of attenuated rising factorial fractional moments (AM) and the method using double Appell sequences (AP), for the parameter estimation of a two-

$r = 2$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
ML	5.97×10^{-8}	0.0006	2.40×10^{-5}	0.0002	0.0076	0.0282
FM	1.17×10^{-8}	0.0262	2.63×10^{-5}	0.0003	8	0.0708
AM	2.89×10^{-6}	0.0006	1.41×10^{-6}	0.0003	0.8273	0.0711
AP	3.18×10^{-7}	0.0036	0.0008	0.0002	10	0.0604

Table 4.27: Estimating a mixture of two geometric distributions with maximum likelihood estimator via the EM algorithm a = 0.1, b = 0.19, p = 0.6, repetition = 10000. Starting values are the true parameters values.

$r = 5$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
ML	2.53×10^{-8}	1.83×10^{-5}	4.51×10^{-8}	3.51×10^{-5}	0.0017	0.0022
FM	3.51×10^{-9}	1.11×10^{-5}	4.80×10^{-7}	4.32×10^{-5}	0.0025	0.0032
AM	1.46×10^{-10}	2.13×10^{-6}	9.83×10^{-10}	4.14×10^{-5}	0.0022	0.0029
AP	5.70×10^{-7}	0.0008	4.29×10^{-5}	6.55×10^{-5}	0.0069	0.0060

Table 4.28: Estimating a mixture of two geometric distributions with maximum likelihood estimator via the EM algorithm a = 0.1, b = 0.19, p = 0.6, repetition = 10000. Starting values are the true parameters values.

component geometric mixture model. We presented our simulation results for each method which allow us to study the robustness of each method for mixtures with different separation between the components and for various sample sizes.

For samples of small sizes, no method appears to show outstanding performance. Although the MLE provides estimates with small variances, we have shown that the ML inferred distributions for small samples have poor goodness of fit. On the other hand, it is quite likely for the FM and AM to provide negative estimates, which are unrealistic because all parameters are probabilities and should lie inside the interval $(0, 1)$. The estimation of b and p using the Appell moments appears to be implausible for samples of small sizes. Therefore, we shall now focus on the estimation results from samples of large sizes and compare these methods. Since the traditional MM is obviously the worst method so we exclude it from our comparison. We draw the estimation results from previous sections, particularly for $n_o = 1000$ and present them in Tables 4.27, 4.28 and 4.29 for $r = 2$, 5 and 10 respectively.

Table 4.27 represents a situation with small separation between the two components ($r = 2$). In terms of the bias, the method of attenuated moments is the most efficient method, except for a; the FM outperforms other methods by providing estimates of a which are closest to the true value. Not surprisingly, the more respectable ML approach provides estimates of all three parameters with the lowest variances. The AM stands out from the other two moment based methods by providing a much lower variance of $\hat{b}$; whereas the AP is the better in controlling the variances of $\hat{a}$ and $\hat{p}$, compared to the other two generalised moment estimators. However, the AP has the largest bias of $\hat{p}$ among all the methods used.

$r = 10$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
ML	3.06×10^{-8}	1.95×10^{-6}	2.90×10^{-8}	2.32×10^{-5}	0.0014	0.0008
FM	1.73×10^{-8}	4.98×10^{-6}	2.78×10^{-8}	2.52×10^{-5}	0.0020	0.0010
AM	8.35×10^{-11}	3.39×10^{-7}	9.41×10^{-12}	2.51×10^{-5}	0.0017	0.0009
AP	3.78×10^{-7}	0.0006	1.90×10^{-5}	4.84×10^{-5}	0.0095	0.0031

Table 4.29: Estimating a mixture of two geometric distributions with maximum likelihood estimator via the EM algorithm a = 0.1, b = 0.19, p = 0.6, repetition = 10000. Starting values are the true parameters values.

For large samples with medium separation ($r = 5$), as seen in Table 4.28, the AM stands out from the others by providing estimates with the minimum bias. In terms of the variance, the MLE is undoubtedly the most efficient method in this case, although the other three methods have variances which are only marginally larger. However, the AP appears to be outperformed by its rivals since it gives estimates with the highest bias and variances.

When both the sample size and the separation are large ($r = 10$ and $n_o = 1000$), as illustrated in Table 4.29, the AM is again the preferred method in terms of bias. With an increase in the magnitude of the separation, all methods provide estimates with considerably small variances; however, the asymptotically most efficient method, the MLE, has the smallest variances for all parameters. Excitingly, the AM estimators have variances which are only marginally larger than the ones given by the MLE; whereas the variances of the FM estimators are only slightly larger than the ones of the AM estimators. The AP remains as the least efficient method in this case.

Figure 4.19, 4.20 and 4.21 show the distribution of the various estimators $\hat{a}$, $\hat{b}$ and $\hat{p}$ respectively, over all replications for a mixture of two geometric distributions with true parameters $a = 0.1$, $b = 0.5$ and $p = 0.6$. The estimation results are drawn from the ones presented in Table 4.28. From these plots, it is obvious that the MLE has the lowest variances, followed by the AM and then the FM. The AP is obviously inferior to the others because it has the largest variances and bias.

To conclude, for a mixture of two geometric distributions, the AM almost always has the lowest bias for all parameters and all sample sizes. MLE remains as the most efficient method in terms of variance in all cases. When the distance between the two components is narrow, the generalised method of moments give poor estimation of b; $Var\left[\hat{b}\right]$ provided by the FM and the AP is significantly larger than the one given by the MLE. However, we observe great improvements in these two methods when the separation between the two components becomes smaller and smaller. *The AM is undoubtedly the most plausible method for a geometric mixture among all of the three moment based methods considered in this chapter. Not only have its estimates the lowest bias, but the variances of the estimators are small and near to the ones given by the all time favourite MLE as the components become closer and closer to each other.*

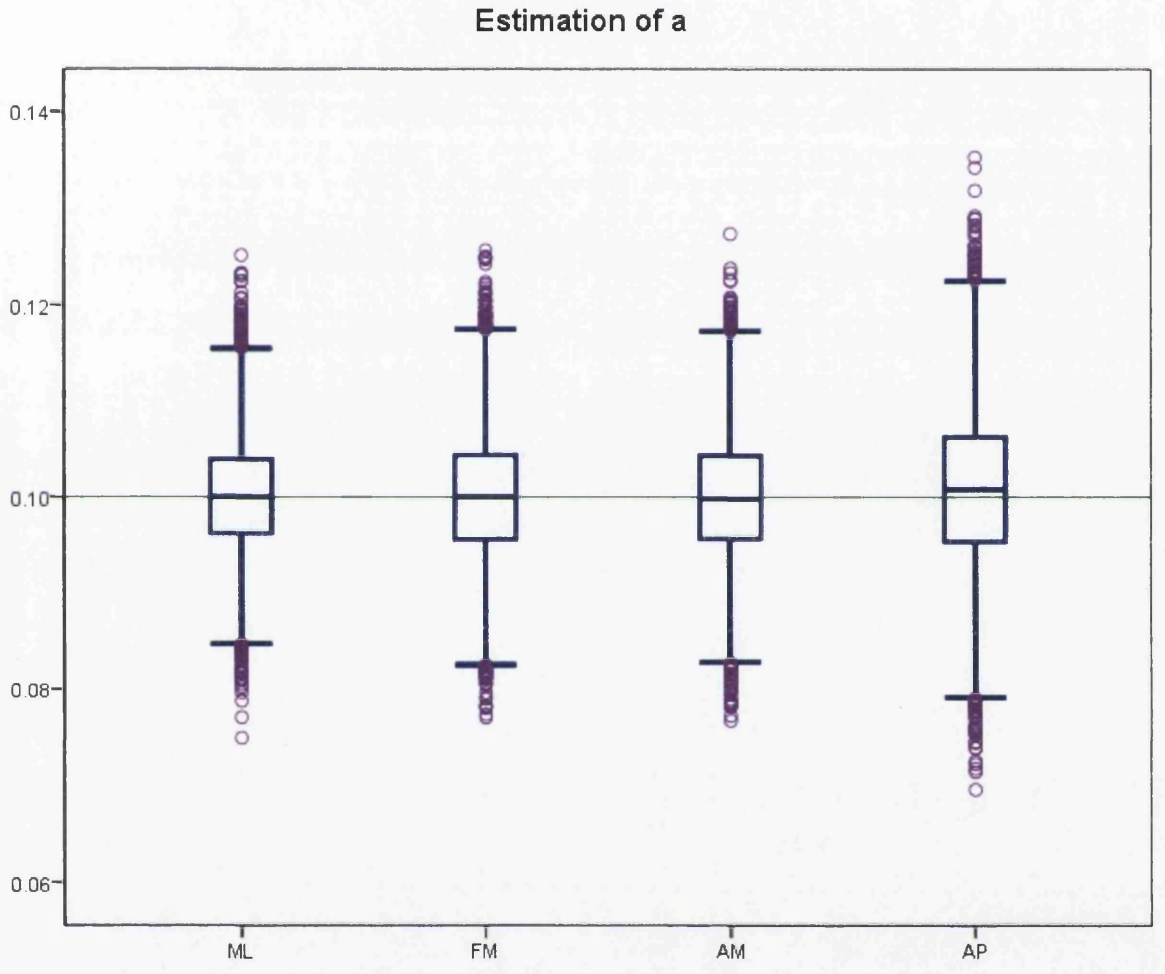


Figure 4.19: Distribution of various estimators $\hat{a}$ for 1000 observations arising from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$.

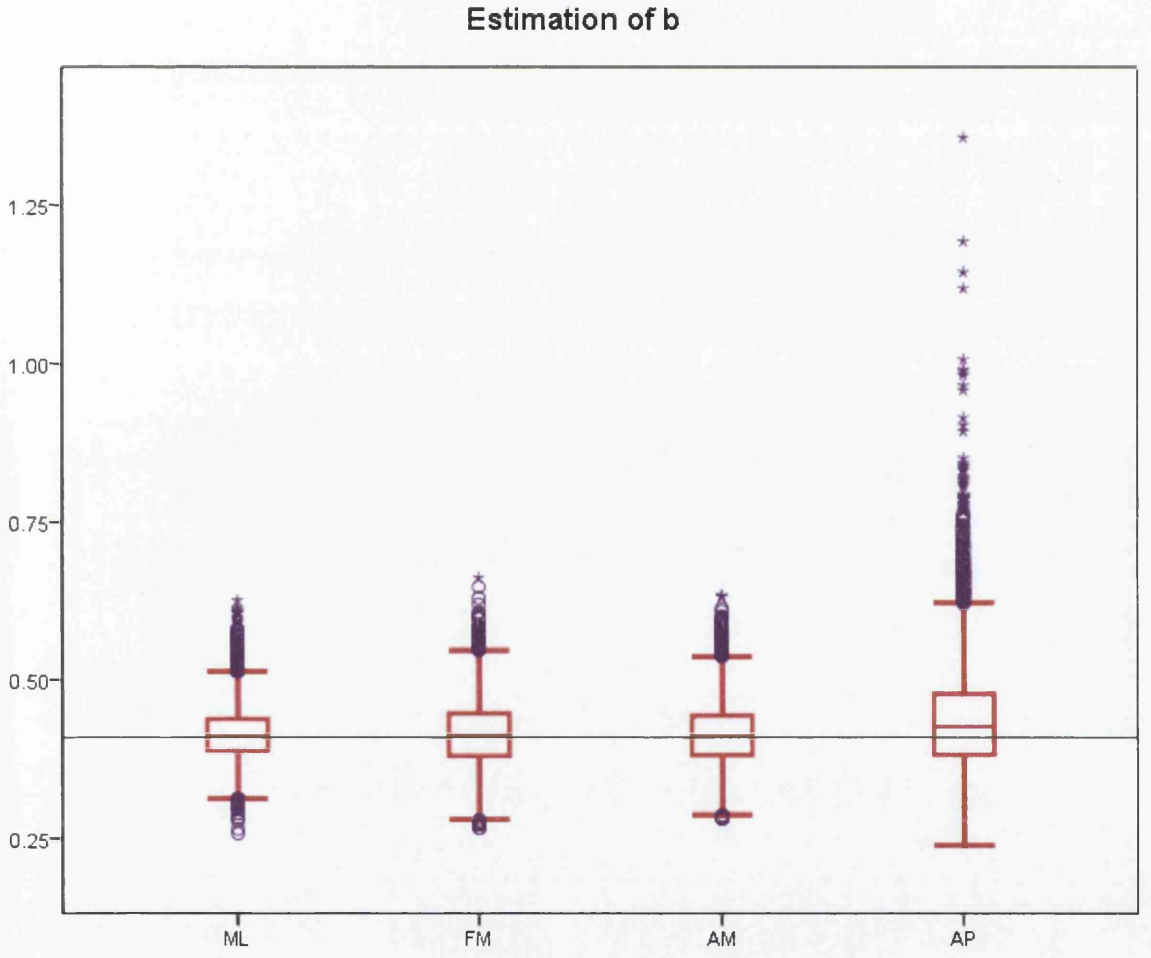


Figure 4.20: Distribution of various estimators $\hat{b}$ for 1000 observations arising from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$.

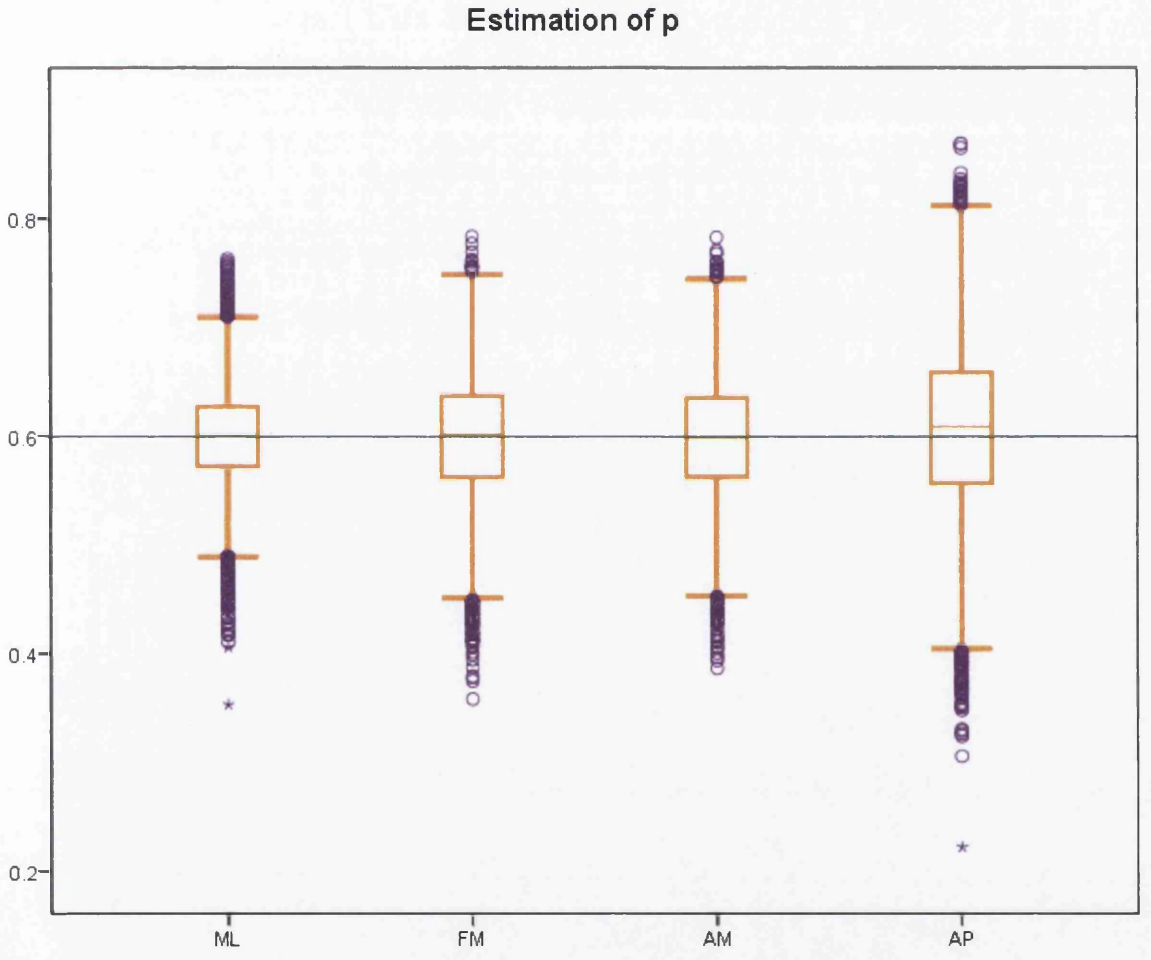


Figure 4.21: Distribution of various estimators $\hat{p}$ for 1000 observations arising from a mixture of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 0.6$.

4.7 Summary

In this chapter, we have studied the application of a few moment based methods to geometric mixtures. Their performances in parameter estimation have been compared with the more standard ML approach. For samples of small sizes, all methods considered here are not very efficient. Our simulation showed that even the MLE inferred distributions have poor goodness of fit for small samples, in spite of the fact that the true parameter values were used to start the iterative process. However, when the number of observations gets larger and larger, all methods provide plausible parameter estimates. Although the MLE is, as expected, the most efficient method in terms of variance, the method of attenuated rising factorial fractional moments stands out to possess much greater efficiency than the other methods which have formal similarities to the moments method. This method provides highly precise estimates with low variances; their estimates have lower bias than the ones given by the MLE when components are well separated and the sample size is sufficiently large.

Chapter 5

Linear Combinations of Distributions

Mixing weights are not necessarily non-negative. For a distribution with a PDF

$$f(t) = \sum_{j=1}^{m} p_j f_j(t),$$

if all p_j's are positive we have a positive mixture distribution; if some of p_j's are negative, it is a linear combination of distributions. In a hidden Markov process in which two states have been clumped into a single level, unless the process is time reversible, the distribution of waiting time in the level may be a linear combination of two exponential distributions with a negative weight.

A search of the literature has not revealed many papers which have investigated the estimation problem of a linear combination of distributions. For such a distribution, it is important that $f(t) \geq 0$ everywhere so that the PDF is valid. Bartholomew (1969) provided two necessary conditions for a mixture of exponential distributions to be a valid PDF, which are

$$\sum_{j=1}^{m} p_j \theta_j \geq 0$$

and

$$p_j > 0$$

and the sufficient conditions for $f(t)$ to be a PDF, which are

$$\sum_{j=1}^{k} p_j \theta_j \geq 0$$

for $k = 1, ..., m$.

The problem of sampling from a positive mixture distribution is straightforward but it

is more complicated if the non-negativity constraints of p_j are to be relaxed. Bignami & de Matteis (1971) discussed this issue and suggested a solution for the problem. We have a different approach for the simulation of a linear combination of distributions and we will explain the solution in a later section.

This chapter will involve a linear combination in which the components are exponential or geometric distributions. All the estimation methods discussed in previous chapters will be applied to the parameter estimation of such distributions and the performance of these estimators will be compared. Our main interest is to suggest reliable and efficient estimation methods for estimating the parameters in a linear combination of two distributions.

5.1 A Linear Combination of Two Exponential Distributions

In this section, we study a number of interesting topics concerned with a linear combination of two exponential distributions. First, we discuss the conditions which are satisfied by the distribution. Simulating a data set arising from a linear combination of distributions is not as easy as the positive mixture; we explain the simulation of this distribution in the second subsection. Lastly, we examine the performance of the estimation methods studied in Chapter 3 in fitting a linear combination of two exponential distributions.

If the distribution of T is a linear combination of the distribution of T_a and the distribution of T_b with weights p and q respectively, where

$$T_a \sim a\exp(-at) \text{ and } T_b \sim b\exp(-bt),$$

then the PDF of T is in the form of

$$f(t_i;\Theta) = pa\exp(-at) + qb\exp(-bt),$$

where $\Theta = (a, b, p)$ and $p + q = 1$. Note that p is positive but not necessarily less than 1. If it is less than 1, T has a mixture of two exponential distributions. However, p can be greater than 1 in the case of waiting time of a strongly time irreversible Markov process. In this case, the weight for T_b, q is negative since $p + q = 1$. The survival function is given by

$$\begin{aligned} S(t) &= p\exp(-at) + q\exp(-bt) \\ &= p\Pr[T_a > t] - |q|\Pr[T_b > t]. \end{aligned}$$

This means that the probability for a person to survive at time t is equivalent to p times the the probability of he/she survives at time t due to reason a, less the probability of he/she survives at time t due to reason b.

5.1.1 Typology of a Linear Combination of Two Exponential Distributions

Let us take a look at the conditions a linear combination of two exponential distributions should satisfy in this section.

Conditions of positivity

1. $a < b$

   The ratio of a and b plays a very important role. $a < b$ means that

   $$r > 1,$$

   The PDF of the linear combination is therefore

   $$f(t) = pa\exp(-at) + qra\exp(-art).$$

2. $p \leq \frac{r}{r-1}$

   The PDF of the linear combination must be non-negative. Hence,

   $$p \geq 0,$$

   and

   $$p + rq \geq 0.$$

   This means that

   $$\begin{aligned} q &\geq -\frac{p}{r} \\ &\Leftrightarrow 1 - p \geq -\frac{p}{r} \\ &\Leftrightarrow p(1 - \frac{1}{r}) \leq 1. \end{aligned}$$

   Hence,

   $$p \leq \frac{r}{r-1}. \tag{5.1}$$

Condition of the mixture

1. $0 \leq p \leq 1$

   This is straightforward. For a mixture of two exponential distributions, p is the probability of T_i comes from the distribution T_a. A probability must be a non-negative value and could not exceed one.

r	$\frac{r^2}{r^2-1}$	$\frac{r}{r-1}$
2	1.3333	2.0000
3	1.1250	1.5000
4	1.0667	1.3333
5	1.0417	1.2500
6	1.0286	1.2000
7	1.0208	1.1667
8	1.0159	1.1429
9	1.0125	1.1250
10	1.0101	1.1111

Table 5.1: Lower and upper bounds for p in a linear combination of two exponential distributions with $a = 0.1$ and $b = 0.1r$.

Condition of linear-non-mixture

1. $1 \leq p \leq \frac{r}{r-1}$

   Since a mixture has $p \leq 1$, for a linear combination to be a non-mixture, we need to have $p \geq 1$. As said before, $p \leq \frac{r}{r-1}$ in order to satisfy the condition that the PDF is always positive. Therefore, we have the above condition.

2. $\frac{r^2}{r^2-1} < p \leq \frac{r}{r-1}$ (non-zero mode)

   Mixtures of exponentials always have a mode at zero. However, general linear combination may have a non-zero mode. In order to have a non-zero mode, we need to have the derivative of the PDF positive at $t = 0$. This means that

$$\frac{\partial}{\partial t}\left[pa\exp(-at) + qb\exp(-bt)\right]_{t=0} > 0$$

$$\begin{aligned} &\Leftrightarrow \quad -a^2p - b^2q > 0 \\ &\Leftrightarrow \quad -p - r^2(1-p) > 0 \\ &\Leftrightarrow \quad p(r^2-1) > r^2 \\ &\Leftrightarrow \quad p > \frac{r^2}{r^2-1}. \end{aligned}$$

   Hence

$$\frac{r^2}{r^2-1} < p \leq \frac{r}{r-1}. \tag{5.2}$$

   For a linear combination of two exponential distributions with a mode, the mixing weight of the first component should have a value between $\frac{r^2}{r^2-1}$ and $\frac{r}{r-1}$.

   Table 5.1 shows the upper bound and the lower bound of p for a linear combination of two exponential distributions with a mode for different ratios of b to a. It is clear

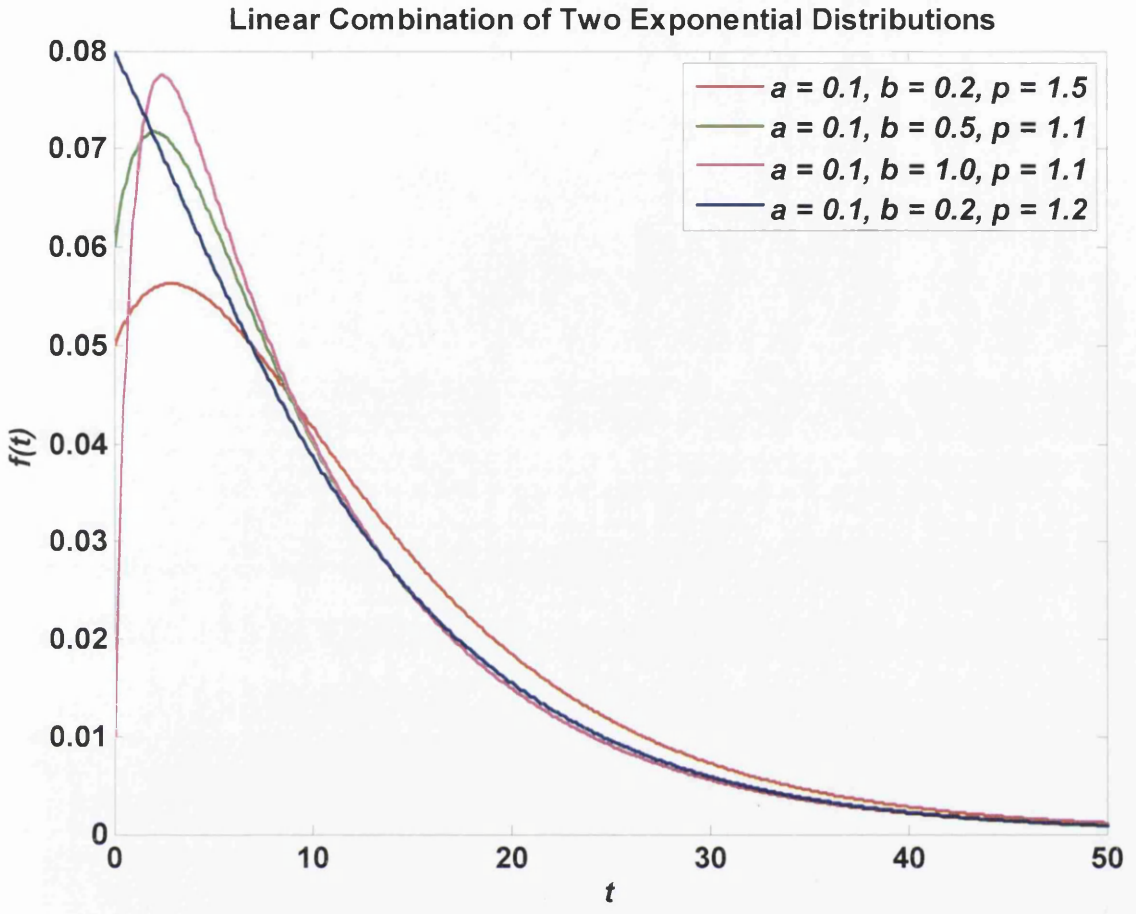


Figure 5.1: PDF plots of linear combinations of two exponential distributions for varying r and p.

from this table that the possible values of p are limited for a linear combination with a large separation between the two distributions.

We see four PDF plots of a linear combination of two exponential distributions for varying r and p in Figure 5.1. When $a = 0.1$, $b = 0.2$ and $p = 1.2$ (indicated by the blue plot), the distribution has no mode because p is less than $\frac{r^2}{r^2-1}$; whereas the other three distributions have modes because their p is between $\frac{r^2}{r^2-1}$ and $\frac{r}{r-1}$.

5.1.2 Simulation of a Linear Combination of Two Exponential Distributions (Non-Mixture)

The simulation of a non-mixture linear combination of two exponential distributions is not as straightforward as a mixture distribution. For a mixture, we can easily generate it with weights p and $q = 1 - p$ of $a\exp(-at)$ and $b\exp(-bt)$, where $0 \leq p \leq 1$. In the case of a

non-mixture, we generate a mixture of $T_a + T_b$ with probability (weight) π_{a+b}, and T_b with probability $1 - \pi_{a+b}$. The relation between π_{a+b} and p is as follows:

$$\pi_{a+b} = p\left(\frac{r-1}{r}\right). \tag{5.3}$$

This being a probability, we need to have $\pi_{a+b} \leq 1$. This inequality follows from the second positivity condition in (5.1).

We would like to know the density of the sum of two random variables X and Y, where

$$\begin{aligned} f_X(x) &= a\exp(-ax), \\ f_Y(y) &= b\exp(-by). \end{aligned}$$

Let $Z = X + Y$ and $f_Z(z)$ its density, so if $z > 0$, then

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(x).f_Y(z-x)\,dx \\ &= \int_0^z a\exp(-ax)\,b\exp(-b(z-x))\,dx \\ &= ab\int_0^z \exp(-bz+(b-a)x)\,dx \\ &= \frac{ab}{b-a}\left[\exp(-bz+(b-a)x)\right]_0^z \\ &= \frac{ab}{b-a}\left[\exp(-bz+bz-az)-\exp(-bz)\right] \\ &= \frac{ab}{b-a}\left[\exp(-az)-\exp(-bz)\right]. \end{aligned}$$

So, the PDF of $T_a + T_b$ (independence of the two is assumed) is

$$\frac{ab}{b-a}\exp(-at) - \frac{ab}{b-a}\exp(-bt), \tag{5.4}$$

and hence the mixture of $T_a + T_b$ and T_b has the PDF

$$f(t) = p\left(\frac{r-1}{r}\right)\left[\frac{ab}{b-a}\exp(-at) - \frac{ab}{b-a}\exp(-bt)\right] + \left(1 - p\left(\frac{r-1}{r}\right)\right)b\exp(-bt)$$

and we know $\dfrac{r-1}{r} = \dfrac{b-a}{b}$, so

$$\begin{aligned} f(t) &= p\left(\frac{b-a}{b}\right)\left[\frac{ab}{b-a}\exp(-at) - \frac{ab}{b-a}\exp(-bt)\right] + \left(1 - p\left(\frac{b-a}{b}\right)\right)b\exp(-bt) \\ &= pa\exp(-at) + (-pa + b - pb + pa)\exp(-bt) \\ &= pa\exp(-at) + (1-p)\,b\exp(-bt) \end{aligned}$$

as required.

r	p	I_{aa}	I_{bb}	I_{pp}	I_{ab}	I_{ap}	I_{bp}
2	1.5	204.3430	6.4381	0.3944	−24.4919	−7.9602	1.4810
		(204.2547)	(6.4445)	(0.3947)	(−24.5145)	(−7.9632)	(1.4824)
5	1.1	119.9060	0.0787	2.2827	−0.7885	−10.3608	0.3520
		(119.7800)	(0.0788)	(2.2826)	(−0.7884)	(−10.3552)	(0.3521)
10	1.1	134.8580	0.1131	13.7654	−1.3349	−23.7608	1.1331
		(135.1203)	(0.1140)	(13.8284)	(−1.3438)	(−23.8359)	(1.1404)

Table 5.2: Theoretical (upper) and simulated (lower) Fisher information for a linear combination of two exponential distributions with fixed $a = 0.1$, and varying r and p.

Remark 3 *In the marginal case where b, i.e. the separation between the two components is one, the PDF is a gamma distribution with shape parameter 2, and the same rate parameter as the exponential distribution.*

Proof. *When $r = 1$, p tends to the upper bound $\frac{r}{r-1}$ and hence π_{a+b}, from (5.3), is one. Therefore, the PDF in this case is given by (5.4), with $a = b$. Since the denominator in (5.4) is now zero, we shall use l'Hôpital's rule to express the PDF in the following form*

$$\begin{aligned} f(t) &= \lim_{b\to a} \frac{ab\exp(-at) - ab\exp(-bt)}{b-a} \\ &= \lim_{b\to a} \frac{\frac{\partial}{\partial b}\left[ab\exp(-at) - ab\exp(-bt)\right]}{\frac{\partial}{\partial b}\left[b-a\right]} \\ &= \lim_{b\to a} \frac{a\exp(-at) - a\exp(-bt) + abt\exp(-bt)}{1} \\ &= a^2 t\exp(-at) \end{aligned}$$

■

Throughout this thesis, we leave aside this marginal case for our study.

5.1.3 Information Matrix and Asymptotic Covariance Matrix of the Maximum Likelihood Estimator

We shall now find the theoretical Fisher information matrix $I(\boldsymbol{\Theta})$ for a linear combination of two exponential distributions using Jalali's (2008) solution, as stated in Section 3.1.6. Following (3.43) to (3.48), we calculate the Fisher information matrices for linear combinations of two exponential distributions with $\boldsymbol{\Theta} = (0.1, 0.2, 1.5)$, $\boldsymbol{\Theta} = (0.1, 0.5, 1.1)$ and $\boldsymbol{\Theta} = (0.1, 1, 1.1)$, particularly when $n_o = 1000$. The theoretical values (upper entries) are shown along with observed values (lower entries) in Table 5.2. Clearly, the conformity between the theory and practice is excellent. Therefore, we find the CRLB of the covariance matrices of estimators for these distributions and show them in Table 5.3. The diagonal elements of the matrices in this table will be used in the last section of this chapter to find the efficiency of all estimators studied by us.

r	p	CRLB of $\boldsymbol{V}[\boldsymbol{\Theta}]$
2	1.5	$\begin{bmatrix} 9.04 \times 10^{-6} & 3.39 \times 10^{-5} & 2.10 \times 10^{-6} \\ 3.39 \times 10^{-5} & 0.0003 & -1.51 \times 10^{-5} \\ 2.10 \times 10^{-6} & -1.51 \times 10^{-5} & 0.0001 \end{bmatrix}$
5	1.1	$\begin{bmatrix} 9.19 \times 10^{-6} & 7.31 \times 10^{-5} & 4.24 \times 10^{-6} \\ 7.31 \times 10^{-5} & 0.0146 & -0.0003 \\ 4.24 \times 10^{-6} & -0.0003 & 6.95 \times 10^{-5} \end{bmatrix}$
10	1.1	$\begin{bmatrix} 8.50 \times 10^{-6} & 8.86 \times 10^{-5} & 1.17 \times 10^{-6} \\ 8.86 \times 10^{-5} & 0.0111 & -0.0001 \\ 1.17 \times 10^{-6} & -0.0001 & 1.33 \times 10^{-5} \end{bmatrix}$

Table 5.3: Cramér-Rao lower bound of $\boldsymbol{V}[\boldsymbol{\Theta}]$ for a linear combination of two exponential distributions with fixed $a = 0.1$ and $n_o = 1,000$, and varying r and p.

5.1.4 Estimation Methods

Since the PDFs are perfectly identical, the estimation methods used to estimate the parameters of a linear combination of exponential distributions is no different to the methods used for a mixture distribution. Therefore, we can easily employ the methods discussed in Chapter 3 for the estimation problem. In this subsection, we show the simulated results using those methods and examine their performances.

Following the procedure described above, we simulate linear combinations of two exponential distributions with three set of parameters, representing three different degrees of separation: $\boldsymbol{\Theta} = (0.1, 0.2, 1.5)$, $\boldsymbol{\Theta} = (0.1, 0.5, 1.1)$ and $\boldsymbol{\Theta} = (0.1, 1, 1.1)$. For each set of parameter, we consider five sample sizes $n_o = (10, 15, 20, 50, 1000)$ to study the behaviour of estimators on samples with different sizes.

The Maximum Likelihood Estimator via the EM Algorithm

First, we consider the MLE. The iterative method used is the EM algorithm where the estimates of a, b and p are updated according to (3.23), (3.24) and (3.20) respectively at each iteration, until the stopping criterion (3.32) is satisfied; we set the tolerance level *tol* as 0.00001.

For a linear combination of two exponential distributions, the value of p during the iterative process is allowed to be greater than 1; when p is updated at a value which does not satisfy condition (5.1), the log-likelihood becomes a complex number. When the true values of the parameters are set as the starting points of the iteration, a majority of the ML estimates diverged from the true values; at some point, the updated value of p did not fulfill (5.1) and caused $\hat{l}^{(k)}$ to be a complex number. In Figure 5.2, we present the

k	$\hat{a}^{(k)}$	$\hat{b}^{(k)}$	$\hat{p}^{(k)}$	$\hat{l}^{(k)}$
20	0.1025	0.2154	1.5574	-3522
21	0.1042	0.2258	1.5816	-3524
22	0.1073	0.2455	1.6282	-3532
23	0.1158	0.2960	1.7568	$-3594-63i$
24	0.1697	0.5153	2.6993	$-3307-299i$
25	-0.4223	0.6246	-2.7737	5552
26	0.0755	0.7420	0.9467	-3553
27	0.077	0.5577	0.9659	-3543

Table 5.4: The ML updated estimates $\Theta^{(k)}$ at each iteration k for an artificial data set, consisting of 1000 observations, simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$. Starting values are set as true values

values of the parameter estimates for a sample arising from a linear combination of two exponential distributions at each iteration. The sample has a size of $n_o = 1000$, and the true parameters are $a = 0.1$, $b = 0.2$ and $p = 1.5$. Using the MLE based on the EM algorithm to fit the data, the starting values of the parameters were set as the true values. We observed that before the sharp turning point at 23^{rd} iteration, the parameter estimates seemed to move in the correct direction. However, at the 23^{rd} iteration, the updated values of the parameters, $\hat{a}^{(23)} = 0.1158$, $\hat{b}^{(23)} = 0.2960$ and $\hat{p}^{(23)} = 1.7568$ led to a complex log-likelihood, $\hat{l}^{(23)} = -3594 - 63i$. In Table 5.4, we show the updated values of the parameters from the 20^{th} to the 27^{th} iteration. After the 24^{th} iteration, as shown in the figure, $\hat{b}^{(k)}$ fell towards $\hat{a}^{(k)}$ as k increased. The iterative process was terminated at the 52^{nd} iteration. The ML estimates are $\hat{a} = 0.0793$, $\hat{b} = 0.0801$, $\hat{p} = 0.9906$ and the log-likelihood is maximised at $\hat{l} = -3534$. We noted the decreasing values of log-likelihood after the 25^{th} iteration; The EM algorithm's property of monotonically convergence in likelihood does not hold here because p and $1 - p$ are not probabilities. This example shows that the EM algorithm is not an ideal tool for the estimation of the parameters in a linear combination of exponential distribution.

This behaviour makes the MLE an unattractive method to estimate a linear combination of two exponential distributions. As we can see from the estimation problem of the single sample discussed here, the final estimates of a and b are 0.0793 and 0.0801 respectively, which are not very distinct from each other. As a consequence, the fitted distribution is reduced to a single exponential distribution with a rate parameter's value being approximately 0.08.

On the same sample, we changed the starting points to $\Theta^{(0)} = (0.1, 0.2, 0.6)$ re-estimated the parameters with the MLE via the EM algorithm. The iteration stopped after 19 iterations with ML estimates $\hat{\Theta} = (0.0791, 0.0797, 0.7005)$ and the log-likelihood $\hat{l} = -3534$. It is worth noting that, regardless of the initial values, the MLE returns similar estimates of a and b, with little difference between the two parameters, and the same log-likelihood. In other words, the MLE identifies the distribution of a sample arising from a linear combination of two exponential distributions as a single exponential distribution. From Figure

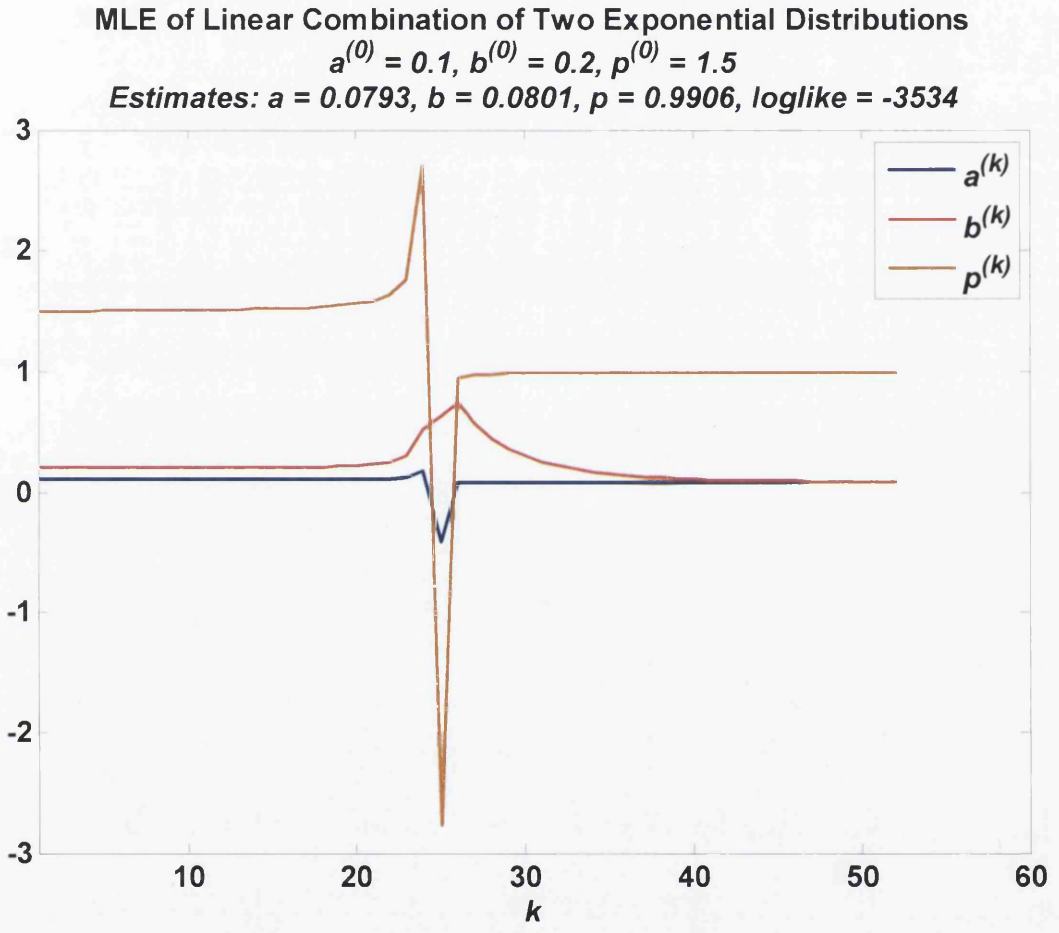


Figure 5.2: The ML updated estimates $\hat{\Theta}^{(k)}$ at each iteration k for an artificial data set, consisting 1000 observations, simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$. Starting values are set as true values.

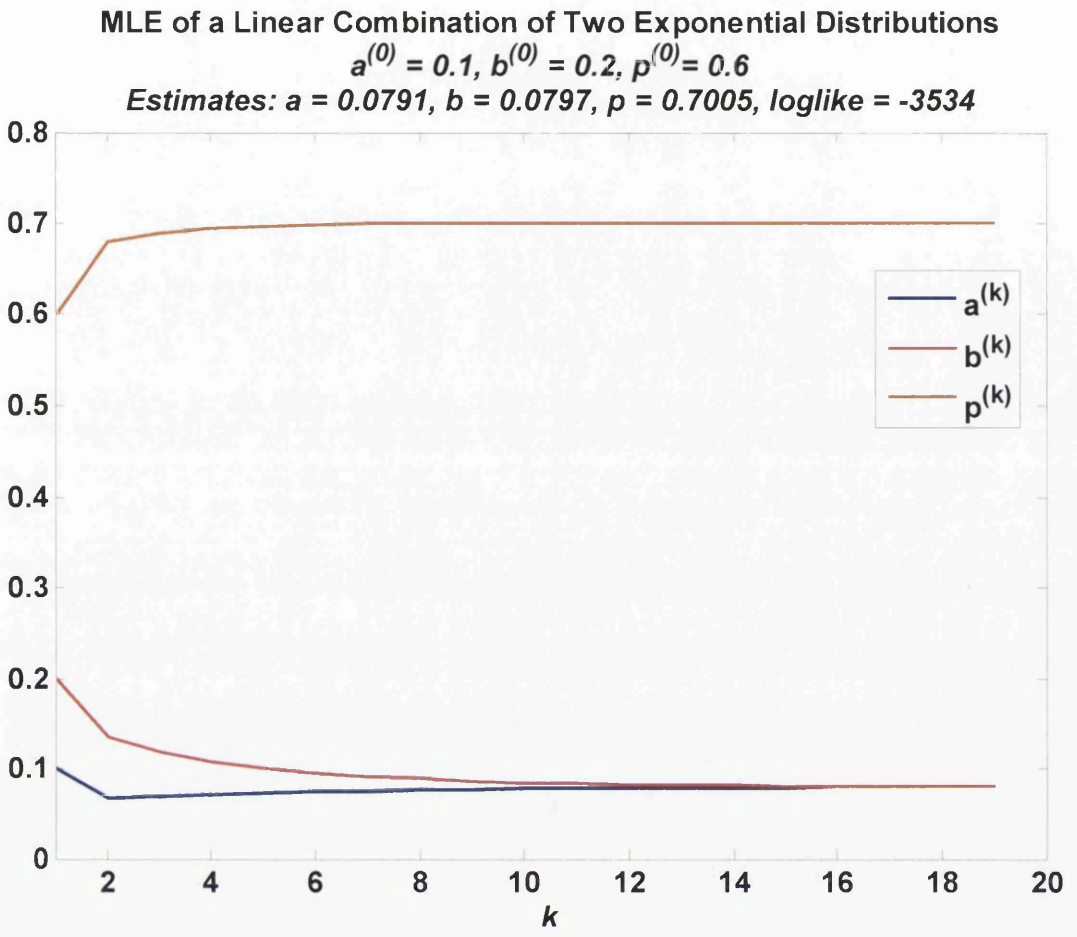


Figure 5.3: The ML updated estimates $\hat{\Theta}^{(k)}$ at each iteration k for an artificial data set, consisting 1000 observations, simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$. Starting values are $a^{(0)} = 0.1$, $b^{(0)} = 0.2$, $p^{(0)} = 0.6$.

5.3, we can see that starting from $\Theta^{(0)} = (0.1, 0.2, 0.6)$, both $\hat{a}^{(k)}$ and $\hat{b}^{(k)}$ converged to a similar point whereas $\hat{p}^{(k)}$ increased monotonically and converged to 0.7005. Figure 5.4 shows the log-likelihood at each iteration; $l^{(k)}$ increased monotonically and terminated at $\hat{l}^{(19)} = -3534$.

It is of our interest to know if the other methods, like the MLE, also recognise the distribution as a single exponential distribution, rather than a linear combination. Therefore, on the same sample, we applied the method of attenuated moments and the method using order statistics to estimate the parameters of the distribution. The estimates from these two methods are indeed more promising than the ML estimates. We used $\kappa = 0.9$ and $c = 0.04$ for the method of attenuated moments, the estimates are $\hat{a} = 0.0927$, $\hat{b} = 0.2465$ and $\hat{p} = 1.2811$, whereas the estimates given by the method using order statistics are $\hat{a} = 0.0943$, $\hat{b} = 0.2366$ and $\hat{p} = 1.3147$.

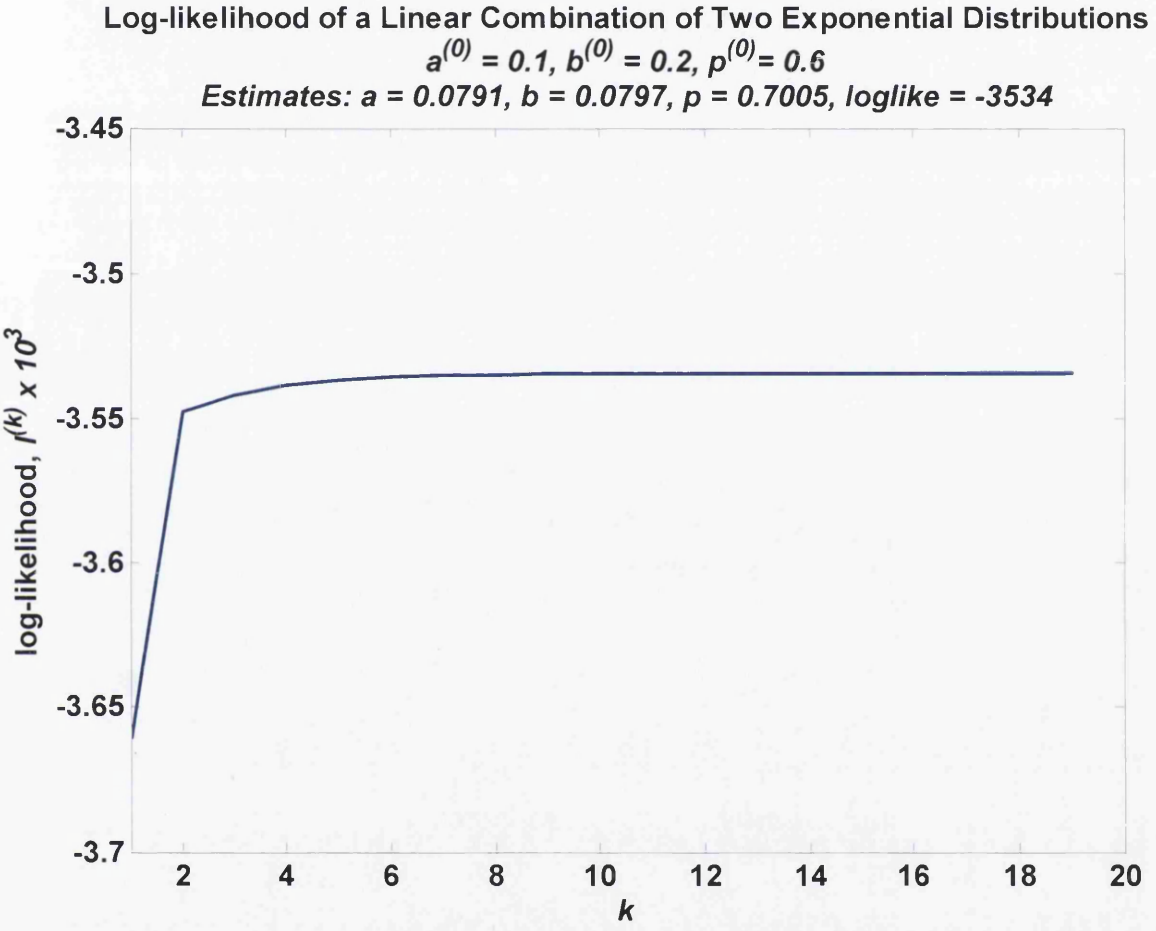


Figure 5.4: The ML updated estimate of log-likelihood $\hat{l}^{(k)}$ at each iteration k for a data set, consisting of 1000 observations, simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$. Starting values are $a^{(0)} = 0.1$, $b^{(0)} = 0.2$, $p^{(0)} = 0.6$.

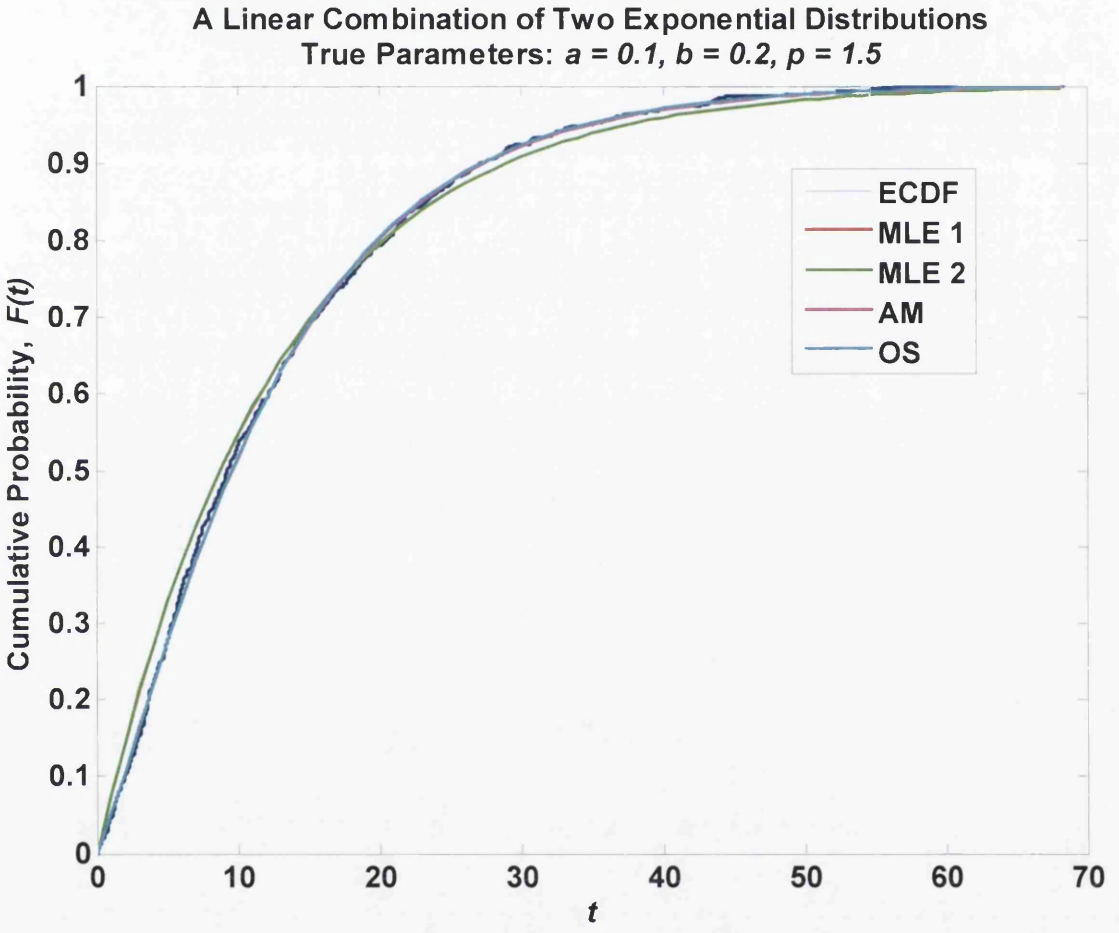


Figure 5.5: Comparison of the ECDF plot of a dataset, consisting of 1000 observations, simulated from a linear combination of two exponential distributions with true parameters $a = 0.1$, $b = 0.2$ and $p = 1.5$, and the fitted CDF plots given by different estimators.

Method	$\hat{a}$	$\hat{b}$	$\hat{p}$	KS	sig
ML 1	0.0791	0.0797	0.7005	0.0637	0.0006
ML 2	0.0793	0.0801	0.9906	0.0638	0.0006
AM	0.0927	0.2465	1.2811	0.0193	0.8476
OS	0.0943	0.2366	1.3147	0.0191	0.8552

Table 5.5: The Kolmogorov-Smirnov test on different estimators of a sample arising from a linear combination of two exponential distributions $a = 0.1$, $b = 0.2$, $p = 1.5$ and $n_o = 1000$.

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.0818	0.0803	0.0794	0.0791	0.0799
$E\left[\hat{b}\right]$	0.2678	0.3794	0.3202	0.0956	0.0806
$E[\hat{p}]$	0.6871	0.6878	0.6890	0.6872	0.6983
$\left(\bar{\hat{a}} - a\right)^2$	0.0003	0.0004	0.0004	0.0004	0.0004
$\left(\bar{\hat{b}} - b\right)^2$	0.0046	0.0322	0.0145	0.0109	0.0143
$\left(\bar{\hat{p}} - p\right)^2$	0.6608	0.6597	0.6577	0.6606	0.6428
$Var[\hat{a}]$	0.0007	0.0005	0.0003	0.0001	4.81×10^{-6}
$Var\left[\hat{b}\right]$	15	176	217	0.0919	5.01×10^{-6}
$Var[\hat{p}]$	0.0106	0.0106	0.0101	0.0081	3.19×10^{-5}
$MSE[\hat{a}]$	0.0010	0.0009	0.0008	0.0006	0.0004
$MSE\left[\hat{b}\right]$	15	176	217	0.1028	0.0143
$MSE[\hat{p}]$	0.6714	0.6703	0.6678	0.6687	0.6428

Table 5.6: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$ for different sample size n_o. Starting values $\boldsymbol{\Theta}^{(0)} = (0.1, 0.2, 0.6)$

We undertook the KS test on each method to examine the goodness of fit and found that, at 5% significance level, we have no evidence to accept that the underlying distribution of the data set is the one suggested by the MLE. *On the other hand, the fitted distribution given by both the attenuated moment estimators and the method using order statistics have relatively shorter KS distances that are highly significant.* (The KS test results are shown in Table 5.5.) In Figure 5.5, we compare the ECDF plot of the data set with the fitted CDF plots given by these estimators. It is clear that both the CDF plots given by the attenuated moment estimator and the method using order statistics are nearer to the ECDF plot, compared to the CDF plot given by the MLE.

In order to compare the performance of the MLE via the EM algorithm with other moment-based estimators, we, as before, consider three degrees of separation $r = 2$, 5 and 10, each with different sample size $n_o = (10, 15, 20, 50, 1000)$. For each case, we simulate 10000 artificial samples and estimate each data set with the MLE via the EM algorithm. We choose the starting values as $a^{(0)} = 0.1$, $b^{(0)} = 0.1r$ and $p^{(0)} = 0.6$, mainly because we are likely to get complex log-likelihood if we set $p^{(0)}$ greater than 1. However, when we set

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.0938	0.0908	0.0902	0.0894	0.0923
$E[\hat{b}]$	1.2538	4	0.9414	0.2641	0.0945
$E[\hat{p}]$	0.8036	0.8006	0.8034	0.8088	0.8445
$(\bar{\hat{a}} - a)^2$	3.81×10^{-5}	8.40×10^{-5}	9.57×10^{-5}	0.0001	5.88×10^{-5}
$(\bar{\hat{b}} - b)^2$	0.5682	11	0.1949	0.0556	0.1644
$(\bar{\hat{p}} - p)^2$	0.0878	0.0897	0.0880	0.0848	0.0653
$Var[\hat{a}]$	0.0012	0.0008	0.0006	0.0003	8.4×10^{-6}
$Var[\hat{b}]$	454	61100	353	4	1.33×10^{-5}
$Var[\hat{p}]$	0.0247	0.0276	0.0276	0.0261	0.0003
$MSE[\hat{a}]$	0.0012	0.0009	0.0007	0.0004	7.00×10^{-5}
$MSE[\hat{b}]$	455	61112	353	4	0.16442
$MSE[\hat{p}]$	0.1125	0.1172	0.1155	0.1109	0.06554

Table 5.7: Performance of the MLE via the EM algorithm for 10,000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 1.1$ for different sample size n_o. Starting values $\Theta^{(0)} = (0.1, 0.5, 0.6)$

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.0937	0.0922	0.0906	0.0898	0.0917
$E[\hat{b}]$	0.6321	0.7137	0.5700	0.1602	0.0937
$E[\hat{p}]$	0.8825	0.8820	0.8841	0.8949	0.9371
$(\bar{\hat{a}} - a)^2$	3.93×10^{-5}	6.13×10^{-5}	8.64×10^{-5}	0.0001	6.88×10^{-5}
$(\bar{\hat{b}} - b)^2$	0.1353	0.0820	0.1849	0.7053	0.8214
$(\bar{\hat{p}} - p)^2$	0.0473	0.0475	0.0466	0.0421	0.0265
$Var[\hat{a}]$	0.0011	0.0007	0.0006	0.0003	7.12×10^{-6}
$Var[\hat{b}]$	39	140	141	3	9.70×10^{-6}
$Var[\hat{p}]$	0.0272	0.0297	0.0302	0.0263	4.14×10^{-5}
$MSE[\hat{a}]$	0.0011	0.0008	0.0007	0.0004	8.00×10^{-5}
$MSE[\hat{b}]$	39	140	141	3	0.8214
$MSE[\hat{p}]$	0.0745	0.0772	0.0768	0.0684	0.0266

Table 5.8: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 1.1$ for different sample size n_o. Starting values $\Theta^{(0)} = (0.1, 1, 0.6)$

$p^{(0)}$ less than 1, this means that $q^{(0)}$ is also less than 1 and is a positive value. We have mentioned before, from (3.16) and (3.20), that the iterative estimates of p and q are always positive if the initial values are set as positive. Therefore, with these initial values, $q^{(k)}$ is never negative. Nevertheless, it is of our interest to investigate how the MLE behaves when the EM algorithm is commenced with these initial values. Tables 5.6 to 5.8 present the estimation results.

Judging from the average values of the estimates $E\left[\hat{\Theta}\right]$ presented in these three tables, we instantly understand that the MLE identifies the distribution as a positive mixture for small samples ($n_o \leq 50$) and it fits a single exponential distribution to large data sets with $n_o = 1000$. In general, the estimation of a is satisfactory with both low bias^2 and variance.

When $r = 2$ and $p = 1.5$, as shown in Table 5.6, the estimator $\hat{b}$ has a lower bias^2 but a larger variance for small n_o, compared to large samples ($n_o = 1000$). It is clear from Table 5.8 that the estimators for $r = 10$ and $p = 1.1$ follow the same behaviour. If we look at the pattern of $E\left[\hat{b}\right]$, it is obvious that, when the sample size increases, $E\left[\hat{b}\right]$ decreases to a value closer to $E[\hat{a}]$. The low variance of estimator $\hat{b}$ when $n_o = 1000$ suggests that most of the 10000 estimators of $\hat{b}$ have similar values to the highly biased $E\left[\hat{b}\right]$. In Table 5.7, we can see that, for data sets with $r = 5$ and $p = 1.1$, the estimation of b is poor for all sample sizes considered in our simulation. When the sample size is small ($n_o \leq 50$), both the bias^2 and variance is large, whereas when the sample size is large ($n_o = 1000$), the behaviour is similar to the cases $r = 2$ and $r = 10$: the estimator $\hat{b}$ has a high bias^2 and a low variance.

At a glance on these three tables, it is clear that the MLE does not provide realistic estimate of $\hat{p}$ in all cases. The reason is straightforward, we started with a positive value of p between 0 and 1, hence $\hat{q}^{(k)}$ is non-negative at each iteration k. This means that the MLE can never fit a linear combination to a data set. However, if we start with a $p^{(0)}$ greater than 1, the iterative process would lead the log-likelihood to a complex form, as shown in Table 5.4, and the EM algorithm would lead $\hat{p}^{(k)}$ back to the region $(0, 1)$. With these reasons in mind, we conclude that the MLE is not an ideal method for estimating a linear combination of two exponential distributions.

Next, we show how an amendment to the MLE could avoid the log-likelihood becoming a complex number during the iterative process. In order to avoid a complex log-likelihood during the iteration process, we need to make sure that, for t_i,

$$pa\exp(-at_i) + (1-p)\,b\exp(-bt_i) \geq 0. \tag{5.5}$$

From the inequality (5.5), we learn that the following two conditions should hold to make sure that we do not have a complex log-likelihood. When $b\exp(-bt_i) - a\exp(-at_i) > 0$, where $t_i > \dfrac{\log r}{(r-1)\,a}$,

$$p < \frac{b\exp(-bt_i)}{b\exp(-bt_i) - a\exp(-at_i)}; \tag{5.6}$$

whereas when $b\exp(-bt_i) - a\exp(-at_i) < 0$, where $t_i < \dfrac{\log r}{(r-1)\,a}$,

$$p > \frac{b\exp(-bt_i)}{b\exp(-bt_i) - a\exp(-at_i)}. \tag{5.7}$$

We know that (5.7) always holds because p is always a positive value. Hence, we only need to be careful about (5.6). Let $b = ra$, then (5.6) becomes

$$p < \frac{(r)\exp(-(r-1)\,at_i)}{(r)\exp(-(r-1)\,at_i) - 1}. \tag{5.8}$$

The denominator should always be positive for any t_i which satisfies the following condition

$$t_i < \frac{\log r}{(r-1)\,a}. \tag{5.9}$$

Prior to the MLE estimation of the data, we should search for any observation which violates (5.9), and group these observations as set C:

$$C = \left\{ t_i : t_i < \frac{\log r}{(r-1)\,a} \right\}. \tag{5.10}$$

The starting point of p is set at a value slightly smaller than

$$\min\left\{ \frac{r\exp(-(r-1)\,at_i)}{r\exp(-(r-1)\,at_i) - 1}; t_i \in C \right\}. \tag{5.11}$$

During the iterative process, if $p^{(k)}$ is greater than

$$\min\left\{ \frac{\left(r^{(k)}\right)\exp\left(-(r-1)^{(k)}\,a^{(k)}t_i\right)}{\left(r^{(k)}\right)\exp\left(-(r-1)^{(k)}\,a^{(k)}t_i\right) - 1}; t_i \in C^{(k)} \right\}, \tag{5.12}$$

we pause the process and exclude any observation t_i which violates (5.8). After the elimination of the so called "bad" data, we continue the estimation using starting point $p^{(k-1)}$. We called this approach the "*shrink sample approach*".

In real life, we will never know if the distribution of a random sample is a positive mixture, so it is likely that we start at the wrong direction by using an unrealistic $p^{(0)}$. In order to avoid making mistakes, we suspect that, if we start from the wrong place, it is very likely that most of the observations would cause violation to the conditions. Hence, we would eliminate most of the data and the sample size would shrink to a small number of observations. For example, if the sample size at the end of estimation is only 50% of the original size, we should restart the estimation by starting from the other side.

In order to investigate whether eliminating any t_i which does not fulfill condition (5.8) will improve the parameter estimation, we simulated a data set, consisting of 100 obser-

vations, arising from a linear combination of two exponential distributions with true parameters $a = 0.1$, $b = 0.2$ and $p = 1.5$ and first fitted the data with the MLE using the normal EM algorithm. The initial values were set as the true values. At the 8^{th} iteration, condition (5.8) was violated and so we observed a complex updated log-likelihood $-345+3i$. The iterative process was terminated at the 24^{th} iteration in which the final estimates are $\hat{a} = 0.0969$, $\hat{b} = 0.1001$, $\hat{p} = 0.5329$ and the log-likelihood is $\hat{l} = -332$. We then re-estimated the parameters using the MLE via the EM algorithm, again we set the initial values as the true values, but this time we paused the process when $p^{(k)}$ was greater than (5.12) and deleted any t_i which violated condition (5.8). We present the updated parameter estimates, for both the normal EM approach and the "*shrink sample approach*", in Table 5.9 (note that $n_o^{(k)}$ denotes the number of observations in the sample at the k^{th} iteration). At the 8^{th} iteration, one observation was eliminated and so the updated log-likelihood did not become a complex number. Similarly, more observations were eliminated from the sample from the 9^{th} to 11^{th} iterations. The iterative process were stopped at the 24^{th} iteration, and the final estimates are $\hat{a} = 0.0859$, $\hat{b} = 0.0862$, $\hat{p} = 1.4555$ and $\hat{l} = -297$. Compared to the normal approach, the "*shrink sample approach*" did improve the parameter estimation. However, the ML estimates of $\hat{a}$ and $\hat{b}$ are still very similar, and hence suggesting a single exponential distribution. It is worth mentioning that the estimation of p has been greatly improved without "bad" data. By shrinking "bad" data from the sample, the MLE via the EM algorithm can at least identify that p is greater than one.

The Method of Fractional Moments

Next, we consider the method of fractional moments in estimating the three parameters of a linear combination of two exponential distributions. Like before, we consider three sets of parameters $\boldsymbol{\Theta} = (0.1, 0.2, 1.5)$, $\boldsymbol{\Theta} = (0.1, 0.5, 1.1)$ and $\boldsymbol{\Theta} = (0.1, 1, 1.1)$, representing three degrees of separation between the two components. We follow the procedure from (3.80) to (3.86) in Section 3.3.1 and consider ten values of fraction $\kappa = (0.1, 0.2, ...1)$. For each κ, we generate 10000 simulated samples of size n_o and estimate each data set with the fractional moment estimator. The best results with minimum measures of errors are shown in Tables 5.10, 5.11 and 5.12.

For a linear combination of two exponential distributions with $r = 2$ and $p = 1.5$, as seen in Table 5.10, the best fraction for estimating a is different for different sample size. However, based on the bias, variance and mean square error, we are certain that any fraction less than or equal to 0.5 should be used to minimise any error in estimating a. For the estimation of b, it seems the best fraction, in terms of both the bias and variance, is $\kappa = 1$ for $n_o \leq 50$. However, our simulation results show that when $\kappa = 1$ is used, it is very likely that the estimate of b is negative; for instance, when $n_o = 50$, 27.26% of the 10000 estimates of b are negative. The number of negative $\hat{b}$ reduces when the sample size increases. When the sample size is as large as $n_o = 1000$, the best fraction is $\kappa = 0.3$,

	Normal EM Approach				Shrink Sample Approach				
k	$\hat{a}^{(k)}$	$\hat{b}^{(k)}$	$\hat{p}^{(k)}$	$\hat{l}^{(k)}$	$\hat{a}^{(k)}$	$\hat{b}^{(k)}$	$\hat{p}^{(k)}$	$\hat{l}^{(k)}$	$n_o^{(k)}$
0	0.1	0.2	1.5	−333	0.1	0.2	1.5	−333	100
1	0.1244	0.2259	1.5885	−330.7131	0.1244	0.2259	1.5885	−330.7131	100
2	0.1248	0.2286	1.5933	−330.723	0.1248	0.2286	1.5933	−330.723	100
3	0.1255	0.2324	1.6002	−330.7437	0.1255	0.2324	1.6002	−330.7437	100
4	0.1265	0.2380	1.6105	−330.7888	0.1265	0.2380	1.6105	−330.7888	100
5	0.128	0.2467	1.6267	−330.8957	0.128	0.2467	1.6267	−330.8957	100
6	0.1305	0.2609	1.6541	−331.1884	0.1305	0.2609	1.6541	−331.1884	100
7	0.1352	0.2870	1.7079	−332.2600	0.1352	0.2870	1.7079	−312.9684	100
8	0.1470	0.3496	1.8530	$-345+3i$	0.147	0.3496	1.8530	−312.6184	99
9	0.4774	1.2377	7.2730	$-387+22i$	0.4916	1.2721	7.5950	−394.9265	94
10	−0.0008	1.3350	−0.0077	−821.0369	0.2144	0.8609	2.7469	−270.4639	89
11	0.0577	0.2383	0.4546	−341.2977	0.3467	0.6683	7.0758	−295.3983	88
12	0.0696	0.1713	0.5074	−334.8381	0.3555	0.5892	8.7246	−294.9981	86
13	0.0762	0.1442	0.5225	−333.2166	0.1299	0.2889	1.9341	−284.6544	86
14	0.0809	0.1296	0.5283	−332.5811	0.1118	0.2083	1.6555	−291.2673	86
15	0.0846	0.1206	0.5307	−332.2773	0.1020	0.1574	1.5369	−294.6770	86
16	0.0875	0.1146	0.5319	−332.1169	0.0960	0.1273	1.4875	−296.2040	86
17	0.0898	0.1104	0.5324	−332.0276	0.0922	0.1097	1.4675	−296.8349	86
18	0.0916	0.1074	0.5326	−331.9764	0.0897	0.0996	1.4598	−297.0789	86
19	0.0931	0.1052	0.5328	−331.9465	0.0881	0.0936	1.457	−297.1679	86
20	0.0942	0.1035	0.5328	−331.929	0.0871	0.0903	1.456	−297.1988	86
21	0.0952	0.1023	0.5329	−331.9187	0.0866	0.0883	1.4557	−297.2092	86
22	0.0959	0.1014	0.5329	−331.9125	0.0862	0.0872	1.4556	−297.2126	86
23	0.0964	0.1007	0.5329	−331.9089	0.0860	0.0866	1.4555	−297.2137	86
24	0.0969	0.1001	0.5329	−331.9067	0.0859	0.0862	1.4555	−297.2140	86

Table 5.9: Comparison of the ML updated estimates $\boldsymbol{\Theta}^{(k)}$ given by the normal EM algorithm and the "shrinked sample approach" at each iteration k for an artificial data set, consisting of 100 observations, simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$. Starting values are set as true values

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	0.0001 (0.1)	0.0002 (0.1)	0.0003 (0.1)	0.0002 (0.4)	4.19×10^{-10} (0.5)
$(\hat{b} - b)^2$	0.0259 (1)	0.0023 (1)	0.0043 (1)	0.0109 (1)	0.0011 (0.4)
$(\hat{p} - p)^2$	0.1073 (0.1)	0.1111 (0.1)	0.1199 (0.1)	0.0377 (0.6)	0.0028 (0.1)
$Var[\hat{a}]$	0.0017 (1)	0.0014 (0.1)	0.0013 (0.1)	0.0010 (0.5)	6.21×10^{-5} (0.3)
$Var[\hat{b}]$	36 (1)	23 (1)	11 (1)	14 (1)	0.0060 (0.3)
$Var[\hat{p}]$	0.9583 (0.2)	0.9872 (0.2)	0.9930 (0.2)	0.6693 (0.2)	0.6564 (0.2)
$MSE[\hat{a}]$	0.0020 (0.1)	0.0016 (0.1)	0.0015 (0.1)	0.0013 (0.5)	6.23×10^{-5} (0.3)
$MSE[\hat{b}]$	36 (1)	23 (1)	11 (1)	14 (1)	0.0072 (0.3)
$MSE[\hat{p}]$	1.0891 (0.2)	1.1268 (0.2)	1.1197 (0.2)	0.8100 (0.2)	0.6614 (0.2)

Table 5.10: Performance of the method of fractional moments for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$ for different sample size n_o.

giving the minimum variance, $Var[\hat{b}] = 0.0060$. The optimal fraction for estimating p is $\kappa = 0.2$. We compare these results to the performance of the fractional moment estimator on the mixture distribution in Table 3.11 and find that, except from $\hat{p}$, the variances of the estimators are lower for a linear combination compared to the positive mixture. The variance of estimator $\hat{p}$ is substantially larger for a linear combination of two exponential distributions because the true mixing weight p is larger than one.

In Table 5.11, we show the estimation results for $r = 5$ and $p = 1.1$. In general, when the κ used is 0.1 or 0.2, $\hat{a}$ is lowly biased. Nevertheless, the variance of $\hat{a}$ is minimised when $\kappa = 0.4$ for large samples of size $n_o = 1000$. It is obvious from the simulation results that the best κ for b is either 0.9 or 1 for any sample size. However, $Var[\hat{b}]$ is generally large, even when the sample size is as large as 1000. It is worth noting that, when $\kappa = 0.1$, the bias^2 of $\hat{p}$ is the lowest for all n_o except when $n_o = 50$. In terms of the mean square error, the best fraction for a and p is any $\kappa \leq 0.5$; whereas we should use a relatively larger fraction ($\kappa = 0.9$ or 1) for b.

From Table 5.12, we can see that, for $r = 10$ and $p = 1.1$, the best fraction for estimating a, in terms of the mean square error, is $\kappa = 0.1$ for all sample sizes we considered. For b, the best fraction is either 0.9 or 1 for $n_o \leq 50$; when $n_o = 1000$, the best κ for b is 0.1. The mean square error of p is minimised when either $\kappa = 0.1$ or $\kappa = 0.2$ is used for all sample sizes.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	5.08×10^{-9} (0.1)	5.90×10^{-6} (0.1)	2.72×10^{-5} (0.1)	6.75×10^{-5} (0.2)	2.51×10^{-10} (1)
$\left(\bar{\hat{b}} - b\right)^2$	0.1425 (1)	0.1590 (1)	0.0795 (0.9)	0.0038 (0.9)	0.1136 (1)
$(\bar{\hat{p}} - p)^2$	0.0003 (0.1)	8.40×10^{-5} (0.1)	0.0002 (0.1)	0.0034 (0.4)	0.0013 (0.1)
$Var[\hat{a}]$	0.0022 (1)	0.0018 (1)	0.0016 (1)	0.0012 (0.1)	3.98×10^{-5} (0.4)
$Var\left[\hat{b}\right]$	111 (1)	35 (1)	26 (1)	21 (0.9)	62 (1)
$Var[\hat{p}]$	0.8316 (0.2)	0.7562 (0.1)	1.0971 (0.5)	0.2915 (0.1)	0.0668 (0.3)
$MSE[\hat{a}]$	0.0025 (1)	0.0021 (0.1)	0.0018 (0.1)	0.0013 (0.1)	4.19×10^{-5} (0.3)
$MSE\left[\hat{b}\right]$	111 (1)	35 (1)	26 (1)	21 (0.9)	62 (1)
$MSE[\hat{p}]$	0.8323 (0.2)	0.7563 (0.1)	1.1152 (0.5)	0.2994 (0.1)	0.0696 (0.3)

Table 5.11: Performance of the method of fractional moments for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 1.1$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	2.24×10^{-7} (0.4)	4.40×10^{-7} (0.4)	6.05×10^{-8} (0.4)	5.17×10^{-11} (0.5)	1.79×10^{-7} (0.1)
$\left(\bar{\hat{b}} - b\right)^2$	0.1438 (0.9)	0.0005 (0.9)	0.1716 (0.9)	0.0080 (0.9)	0.0033 (0.1)
$(\bar{\hat{p}} - p)^2$	4.28×10^{-5} (0.5)	0.0008 (0.6)	0.0002 (0.7)	0.0019 (0.8)	3.04×10^{-5} (0.1)
$Var[\hat{a}]$	0.0024 (1)	0.0019 (0.1)	0.0015 (0.1)	0.0006 (0.1)	1.45×10^{-5} (0.1)
$Var\left[\hat{b}\right]$	300 (1)	92 (1)	16 (1)	287 (0.9)	0.1187 (0.1)
$Var[\hat{p}]$	1.0944 (0.2)	1.1898 (0.1)	0.7754 (0.1)	0.4052 (0.2)	0.0011 (0.1)
$MSE[\hat{a}]$	0.0027 (0.1)	0.0019 (0.1)	0.0015 (0.1)	0.0006 (0.1)	1.47×10^{-5} (0.1)
$MSE\left[\hat{b}\right]$	301 (1)	93 (1)	17 (1)	287 (0.9)	0.1219 (0.1)
$MSE[\hat{p}]$	1.1057 (0.2)	1.2037 (0.1)	0.7895 (0.1)	0.4157 (0.2)	0.0011 (0.1)

Table 5.12: Performance of the method of fractional moments for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 1.1$ for different sample size n_o.

r	p	Theoretical Optimal κ	Theoretical $Var[\hat{a}]$	Practical Optimal κ	Practical $Var[\hat{a}]$
2	1.5	0.4224	9.98×10^{-5} (6.38×10^{-5})	0.3	6.21×10^{-5}
5	1.1	0.2273	2.37×10^{-5} (4.42×10^{-5})	0.3	3.98×10^{-5}
10	1.1	0.0753	1.39×10^{-5} (1.52×10^{-5})	0.1	1.45×10^{-5}

Table 5.13: Theoretical and simulated minimum variance of fractional moment estimator $\hat{a}$ given by the optimal κ for a linear combination of two exponential distributions with fixed $a = 0.1$, $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

r	p	Theoretical Optimal κ	Theoretical $Var[\hat{b}]$	Practical Optimal κ	Practical $Var[\hat{b}]$
2	1.5	0.4350	0.0051 (0.0072)	0.3	0.0060
5	1.1	0.2465	0.0840 (4.3665)	0.3	0.0668
10	1.1	0.0609	0.0821 (0.1028)	0.1	0.1187

Table 5.14: Theoretical and simulated minimum variance of fractional moment estimator $\hat{b}$ given by the optimal κ for a linear combination of two exponential distributions with fixed $a = 0.1$, $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

The reason we see the best fraction for estimating b is $\kappa = 1$ for samples of small sizes has been explained in Section 3.3.5: b is more sensitive to y when κ is a fraction (see Figure 3.17). Compared to a sample of large size, a sample of small size is more likely to have an estimate of y with a larger deviation from the true value. When fractional moments are used, a small deviation of $\hat{y}$ from the true value is highly likely to increase the estimate of b and result in an over-estimation of b. This is why the variance of estimator $\hat{b}$ is larger when fractional moments are used, compared to ordinary moments.

For a linear combination of two exponential distributions, we discovered that the agreement between the theoretical variances of the estimators and the practical variances of the estimators is not as strong as for the positive mixture case. In Tables 5.13, 5.14 and 5.15, we focus on large data sets with $n_o = 1000$ and compare the theoretical variances of fractional estimators approximated by (3.87) with the practical values obtained from our simulation experiment. We used the optimal κ suggested by the theory to fit another 10000 simulated data sets and recorded the observed variances of the estimators in brackets underneath the approximated theoretical values in these tables. For a, as seen in Table 5.13, the $Var[\hat{a}]$ returned by the suggested κ is close to, but still marginally larger than the one we obtained from our simulation experiments. This is similar for b, as shown in Table 5.14, except when $r = 5$. In Table 5.15, it is clear that (3.87) under-estimates the variance of $\hat{p}$. We are not

r	p	Theoretical Optimal κ	Theoretical $Var[\hat{p}]$	Practical Optimal κ	Practical $Var[\hat{p}]$
2	1.5	0.4278	0.2093 (1.6518)	0.2	0.6564
5	1.1	0.2298	0.0041 (0.1199)	0.3	0.0668
10	1.1	0.0753	1.39×10^{-5} (0.0012)	0.1	0.0011

Table 5.15: Theoretical and simulated minimum variance of fractional moment estimator $\hat{p}$ given by the optimal κ for a linear combination of two exponential distributions with fixed $a = 0.1$, $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

surprised to see the lack of agreement between the theory and practice, because (3.87) only provides an approximation to the variance of estimator. The fact that q is negative may have affected the accuracy of the approximation.

Since the suggested κ is able to provide estimates with low variances which are close to the ones we obtained in our simulation experiments, we are certain that a small fraction κ with any value between 0.1 and 0.3 will return estimates of a, b and p with both low bias and low variance, especially when the sample size is large enough. From Tables 5.13 to 5.15, we also notice that the best fraction κ decreases with r.

In Chapter 3, we suggested a way for users to verify that a good κ has been chosen for the estimation problem. We shall now demonstrate that this approach can also be taken when the true value of p is greater than one. We simulated a sample, of size 1000, from a linear combination of two exponential distributions with true parameters $a = 0.1$, $b = 1$ and $p = 1.1$, and estimated the parameters from the data set using ten different κ ranging from 0.1 to 1. Therefore, we have ten sets of parameter estimates. Which set is closest to the true values? For such a distribution, we know, from Tables 5.13 to 5.15, that the best κ should be 0.1. Since we do not know this in practice, we substituted these sets of estimates into (3.87) and plot the resulted $Var\left[\hat{\Theta}\right]$ versus κ alongside the theoretical variances given by the true values in Figure 5.6 for a, b and p respectively. We can confirm that the best κ (in this case is 0.1) does provide estimates that make $Var\left[\hat{\Theta}\right]$ smallest when they are substituted into (3.87). Therefore, in practice, users should simply fit a raw sample using a number of different κ and substitute the yielded sets of estimates into (3.87). The most precise set of estimates is then the one which has the smallest $Var\left[\hat{\Theta}\right]$.

The Method of Attenuated Moments

We have seen in Chapter 3 that the method of attenuated moments produces parameter estimates for a mixture of two exponential distributions with relatively higher precision, compared to the method of fractional moments. We have shown earlier that the attenuated moment estimator does a better job than the MLE in fitting a linear combination of two exponential distributions to a simulated data set.

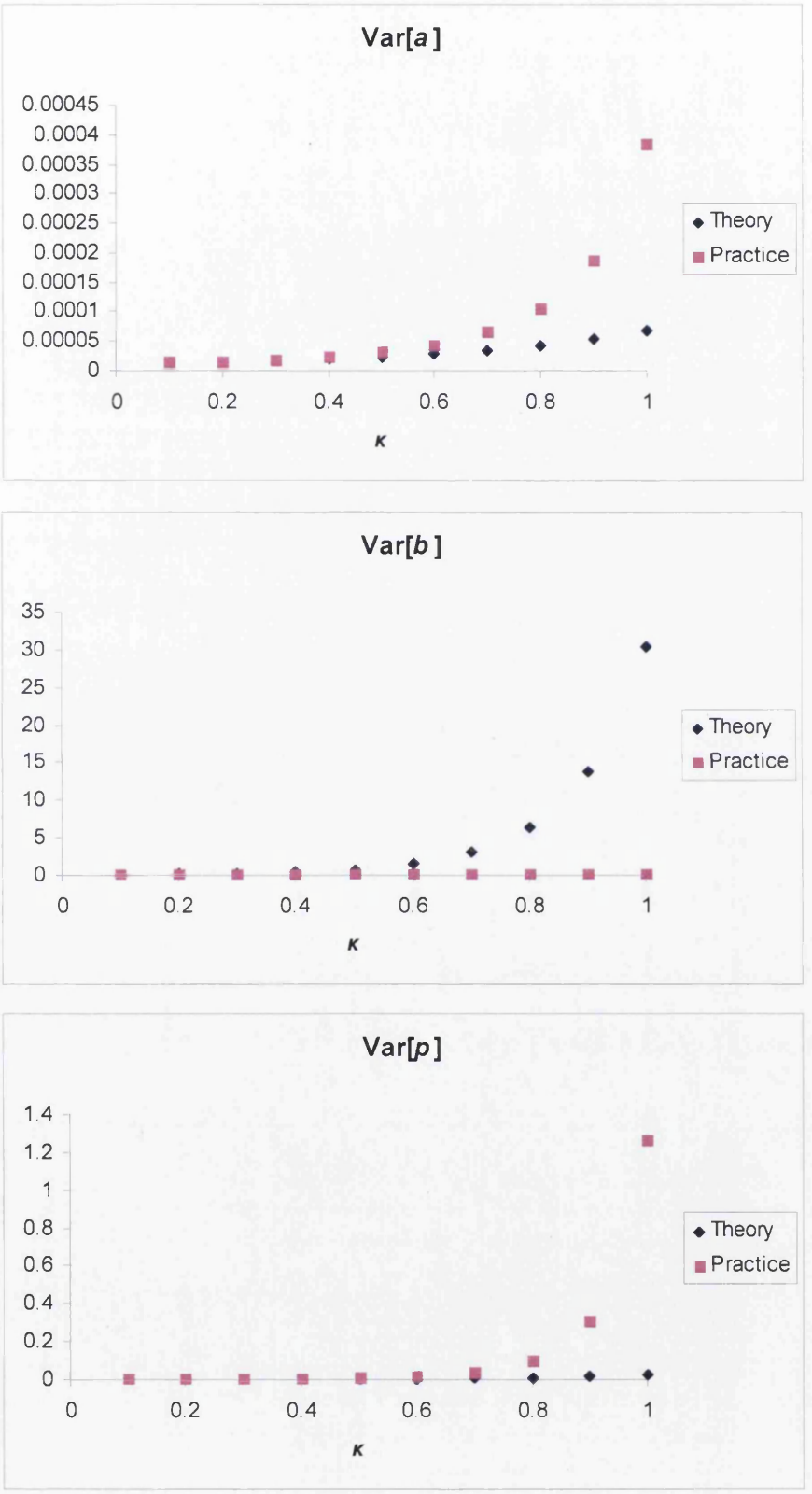


Figure 5.6: Asymptotic variance of the fractional moment estimator given by true parameters and parameter estimates versus κ, based on a data set, consisting of 1000 observations, simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 1.1$.

Here, we show the estimation results for linear combination of two exponential distributions with different separation and different mixing weights. The procedures we took were the ones from (3.108) to (3.115). Once again, we study the performance of this method on a linear combination of two exponential distributions with three different separation between the two populations, $r = (2, 5, 10)$. Like before, we consider ten values of fraction, $\kappa = (0.1, 0.2, ..., 1)$ while for each fraction we consider nineteen values of attenuation, $c = (0.01, 0.02, ..., 0.1)$. For each of these 100 combinations of κ and c, we simulate 10000 artificial samples with the specified parameters and estimate them with the attenuated moment estimator. We then draw conclusions on the best combination that returns the minimum measure of errors.

For a linear combination of two exponential distributions with $r = 2$ and $p = 1.5$, as seen in Table 5.16, one should use $\kappa = 1$ and $c = 0.01$ for samples with sizes $n_o \leq 50$ to minimise $Var[\hat{a}]$; when $n_o = 1000$, the optimal combination of κ and c is $(0.9, 0.04)$ for a, with $Var[\hat{a}] = 6.17 \times 10^{-5}$. With the same sample size, the best combination of κ and c for b is $(0.6, 0.04)$, which gives the minimum $Var\left[\hat{b}\right] = 0.0052$. Most of the best κ for p are 0.1 with an attenuation c between 0.02 and 0.1. Compared to the fractional moment estimators, the variances of the attenuated moment estimators ($\hat{a}, \hat{b}$ and $\hat{p}$) are greatly reduced. Judging from the mean square error, in general, we should use a large fraction ($0.5 \leq \kappa \leq 1$) and a small attenuation ($0.01 \leq c \leq 0.05$) for the estimation of a and b; whereas for p, we should use a small fraction ($0.1 \leq \kappa \leq 0.2$) and an attenuation ranging from 0.02 to 0.1.

Now, we study the performance of the attenuated moment estimator for $r = 5$ and $p = 1.1$ in Table 5.17. Similar to $r = 2$, the ideal combination of κ and c for the estimation of a is $(1, 0.01)$ for small sample size ($n_o \leq 50$); whereas when $n_o = 1000$, the minimum $Var[\hat{a}] = 3.29 \times 10^{-5}$ is given by $\kappa = 0.5$ and $c = 0.02$. For b, when the number of observations in a sample is smaller than or equal to 50, one should use a large fraction ($\kappa = 0.9$ or 1) with an attenuation c between 0.01 and 0.08. We cannot make a conclusion on the best combination of κ and c for p; when $n_o = 1000$, the ideal combination for p is $(0.6, 0.06)$ which minimises both the variance and the mean square error.

As seen in Table 5.18, for $r = 10$ and $p = 1.1$, the ideal combination of κ and c for a is $(1, 0.01)$ for $n_o \leq 20$, $(0.2, 0.01)$ for $n_o = 50$, and $(0.3, 0.02)$ for $n_o = 1000$. For b, the best κ for $n_o \leq 50$ is either 0.9 or 1 with c ranging from 0.01 to 0.09; whereas when $n_o = 1000$, the best combination of κ and c for b is $(0.1, 0.04)$ in terms of the mean square error. For p, the best combination of κ and c is $(0.2, 0.04)$ when $n_o = 1000$, where the minimum $Var[\hat{p}]$ is 0.0008.

In general, regardless of the separation between the two distributions, the best combination for the estimation of a and b is a large fraction and a low attenuation when the sample size is small. On the other hand, when n_o is as large as 1000, the best combination for these two parameters is a low fraction with a low attenuation. For the estimation of p, in most cases, the best combination is a small fraction and a small attenuation, regardless of the sample size.

$r=2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}}-a)^2$	5.46×10^{-5} $(0.9,0.01)$	7.14×10^{-5} $(0.9,0.01)$	6.89×10^{-5} $(0.9,0.01)$	3.20×10^{-5} $(0.8,0.01)$	4.88×10^{-13} $(0.6,0.02)$
$\left(\bar{\hat{b}}-b\right)^2$	0.0009 $(1,0.08)$	0.0013 $(1,0.06)$	2.49×10^{-5} $(1,0.01)$	0.0002 $(1,0.1)$	0.0009 $(0.6,0.02)$
$(\bar{\hat{p}}-p)^2$	2.03×10^{-6} $(0.8.0.03)$	5.12×10^{-6} $(0.9,0.03)$	3.81×10^{-6} $(0.8,0.02)$	8.42×10^{-9} $(0.9,0.04)$	3.88×10^{-5} $(0.2,0.02)$
$Var[\hat{a}]$	0.0008 $(1,0.01)$	0.0008 $(1,0.01)$	0.0008 $(1,0.01)$	0.0006 $(1,0.01)$	6.17×10^{-5} $(0.9,0.04)$
$Var\left[\hat{b}\right]$	7 $(0.8,0.02)$	17 $(0.8,0.02)$	11 $(0.9,0.04)$	30 $(1,0.05)$	0.0052 $(0.6,0.04)$
$Var[\hat{p}]$	0.8323 $(0.1,0.02)$	0.7165 $(0.2,0.08)$	0.6070 $(0.1,0.09)$	0.3755 $(0.1,0.07)$	0.3907 $(0.1,0.1)$
$MSE[\hat{a}]$	0.0009 $(1,0.01)$	0.0009 $(1,0.01)$	0.0009 $(1,0.01)$	0.0006 $(0.8,0.01)$	6.17×10^{-5} $(0.9,0.04)$
$MSE\left[\hat{b}\right]$	7 $(0.8,0.02)$	17 $(0.8,0.02)$	11 $(0.9,0.04)$	30 $(1,0.05)$	0.0063 $(0.5,0.02)$
$MSE[\hat{p}]$	0.9528 $(0.1,0.02)$	0.8916 $(0.2,0.08)$	0.8379 $(0.1,0.09)$	0.6058 $(0.1,0.07)$	0.7204 $(0.1,0.1)$

Table 5.16: Performance of the method of attenuated fractional moments for 10000 data sets simulated from a linear combination of two exponential distributions with $a=0.1$, $b=0.2$ and $p=1.5$ for different sample size n_o.

$r=5$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}}-a)^2$	4.55×10^{-7} $(0.1,0.01)$	1.86×10^{-5} $(0.1,0.01)$	2.86×10^{-5} $(0.3,0.01)$	3.38×10^{-5} $(0.3,0.01)$	2.97×10^{-11} $(0.2,0.04)$
$\left(\bar{\hat{b}}-b\right)^2$	0.0558 $(1,0.01)$	0.0086 $(1,0.08)$	2.52×10^{-5} $(1,0.09)$	0.0005 $(1,0.08)$	0.0003 $(1,0.04)$
$(\bar{\hat{p}}-p)^2$	9.00×10^{-7} $(0.2,0.01)$	3.27×10^{-6} $(0.9,0.02)$	1.14×10^{-6} $(0.8,0.02)$	1.48×10^{-8} $(0.6,0.01)$	1.24×10^{-5} $(0.2,0.08)$
$Var[\hat{a}]$	0.0010 $(1,0.01)$	0.0010 $(1,0.01)$	0.0011 $(1,0.01)$	0.0010 $(1,0.01)$	3.29×10^{-5} $(0.5,0.02)$
$Var\left[\hat{b}\right]$	54 $(0.9,0.05)$	94 $(1,0.08)$	104 $(1,0.01)$	159 $(1,0.03)$	0.4234 $(0.3,0.05)$
$Var[\hat{p}]$	0.5366 $(1,0.03)$	0.4350 $(0.8,0.02)$	0.4476 $(1,0.04)$	0.3677 $(0.1,0.03)$	0.0413 $(0.6,0.06)$
$MSE[\hat{a}]$	0.0011 $(1,0.01)$	0.0011 $(1,0.01)$	0.0011 $(1,0.01)$	0.0011 $(1,0.01)$	3.48×10^{-5} $(0.5,0.02)$
$MSE\left[\hat{b}\right]$	54 $(0.9,0.05)$	94 $(1,0.08)$	104 $(1,0.01)$	159 $(1,0.03)$	0.4467 $(0.3,0.05)$
$MSE[\hat{p}]$	0.5418 $(0.3,0.02)$	0.4363 $(0.8,0.02)$	0.4560 $(1,0.04)$	0.3764 $(0.1,0.03)$	0.0433 $(0.6,0.06)$

Table 5.17: Performance of the method of attenuated fractional moments for 10000 data sets simulated from a linear combination of two exponential distributions with $a=0.1$, $b=0.5$ and $p=1.1$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	8.30×10^{-10} (0.4, 0.09)	1.54×10^{-8} (0.4, 0.09)	4.68×10^{-10} (0.4, 0.02)	3.98×10^{-10} (0.8, 0.05)	1.25×10^{-7} (0.2, 0.04)
$(\bar{\hat{b}} - b)^2$	0.0038 (1, 0.04)	0.0275 (1, 0.06)	0.1289 (1, 0.07)	0.0245 (1, 0.03)	0.0008 (1, 0.09)
$(\bar{\hat{p}} - p)^2$	4.86×10^{-7} (0.7, 0.05)	2.30×10^{-8} (0.6, 0.08)	6.93×10^{-9} (1, 0.07)	5.38×10^{-7} (1, 0.03)	1.74×10^{-5} (0.2, 0.04)
$Var[\hat{a}]$	0.0010 (1, 0.01)	0.0011 (1, 0.01)	0.0010 (1, 0.01)	0.0006 (0.2, 0.01)	1.42×10^{-5} (0.3, 0.02)
$Var[\hat{b}]$	80 (1, 0.01)	268 (1, 0.03)	360 (0.9, 0.05)	194 (0.9, 0.09)	0.0814 (0.1, 0.06)
$Var[\hat{p}]$	0.4178 (1, 0.01)	0.4153 (0.3, 0.08)	0.4621 (0.1, 0.02)	0.1644 (0.1, 0.02)	0.0008 (0.2, 0.04)
$MSE[\hat{a}]$	0.0010 (1, 0.01)	0.0011 (1, 0.01)	0.0010 (1, 0.01)	0.0006 (0.2, 0.01)	1.43×10^{-5} (0.3, 0.02)
$MSE[\hat{b}]$	81 (1, 0.01)	269 (1, 0.03)	362 (0.9, 0.05)	195 (0.9, 0.09)	0.0825 (0.1, 0.04)
$MSE[\hat{p}]$	0.4181 (1, 0.01)	0.4191 (0.3, 0.08)	0.4686 (0.1, 0.02)	0.1669 (0.1, 0.02)	0.0009 (0.2, 0.04)

Table 5.18: Performance of the method of attenuated fractional moments for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 1.1$ for different sample size n_o.

From these three tables, we observe that the bias^2 are very small for all r, especially for the estimators $\hat{a}$ and $\hat{p}$. This means that, compared to the MLE, the attenuated moment estimator is better in "recognising" a data set arising from a linear combination of two distributions. Previously, we have learned that the MLEs are highly biased and it fails to return a mixing weight with a value greater than 1. Based on this argument, the method of attenuated moments outperforms the MLE in the estimation problem of a linear combination of two exponential distributions.

Tables 5.19, 5.20 and 5.21 compare the approximated theoretical variance of attenuated moment estimators given by (3.119) with the practical values obtained from the simulation experiment; we focus on large data sets with $n_o = 1000$. Since the optimal combinations of κ and c suggested by the theory differ from the ones observed from our simulation experiments, we used the suggested combination to estimate from another 10000 data sets and the observed variances of the estimators are presented in brackets underneath the theoretical values in the table. For a, as seen in Table 5.19, the variances of the estimators given by the suggested combination are indeed close to, but marginally larger than the ones we obtained previously. For b, as in Table 5.20, the agreement between the theoretical and observed combination of κ and c are indeed quite good for $r = 2$ and 10. The practical variance of the estimator $\hat{b}$ is slightly larger than the one we observed from our previous simulation experiment. For p, like the fractional moment estimator, the variance of $\hat{p}$ is under-estimated

r	p	Theoretical Optimal κ and c	Theoretical $Var[\hat{a}]$	Practical Optimal κ and c	Practical $Var[\hat{a}]$
2	1.5	$(0.6156, 0.0162)$	9.55×10^{-5} (6.35×10^{-5})	$(0.9, 0.04)$	6.17×10^{-5}
5	1.1	$(0.3734, 0.0203)$	2.24×10^{-5} (3.75×10^{-5})	$(0.5, 0.02)$	3.29×10^{-5}
10	1.1	$(0.3908, 0.0579)$	1.64×10^{-5} (1.69×10^{-5})	$(0.3, 0.02)$	1.42×10^{-5}

Table 5.19: Theoretical and simulated minimum variance of attenuated moment estimator $\hat{a}$ given by the optimal combination of κ and c for a linear combination of two exponential distributions with fixed $a = 0.1$, $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

r	p	Theoretical Optimal κ and c	Theoretical $Var[\hat{b}]$	Practical Optimal κ and c	Practical $Var[\hat{b}]$
2	1.5	$(0.6291, 0.0195)$	0.0048 (0.0073)	$(0.6, 0.04)$	0.0052
5	1.1	$(0.4842, 0.0525)$	0.0684 (1.9702)	$(0.3, 0.05)$	0.4234
10	1.1	$(0.1854, 0.0640)$	0.0690 (0.0888)	$(0.1, 0.06)$	0.0814

Table 5.20: Theoretical and simulated minimum variance of attenuated moment estimator $\hat{b}$ given by the optimal combination of κ and c for a linear combination of two exponential distributions with fixed $a = 0.1$, $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

r	p	Theoretical Optimal κ and c	Theoretical $Var[\hat{p}]$	Practical Optimal κ and c	Practical $Var[\hat{p}]$
2	1.5	$(0.6221, 0.0178)$	0.1977 (1.3720)	$(0.1, 0.1)$	0.3907
5	1.1	$(0.4284, 0.0380)$	0.0036 (0.0730)	$(0.6, 0.06)$	0.0413
10	1.1	$(0.3261, 0.1218)$	0.0007 (0.00120)	$(0.2, 0.04)$	0.0008

Table 5.21: Theoretical and simulated minimum variance of attenuated moment estimator $\hat{p}$ given by the optimal combination of κ and c for a linear combination of two exponential distributions with fixed $a = 0.1$, $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

by (3.119), as seen in Table 5.21. The fact that one of the mixing weights, q is negative might have created some errors in the calculation of the approximated theoretical variance of estimator.

By now, the trend is clear, the best κ decreases when the separation between the two distributions increases. It is clear that we can use a wide range of κ and c and still be able to obtain reasonable estimates of the parameters, since the variances of the estimators given by any of these combinations are generally low.

In Chapter 3, we demonstrated how users can decide on the best set of estimates after estimating the parameters from a raw sample with a few combinations of κ and c. Our suggestion is, the ones that give the lowest $Var\left[\hat{\Theta}\right]$ in (3.119) should be chosen. To check if we can do the same when the underlying distribution is a linear combination, we estimated the same data set simulated in Figure 5.6, where the true parameters are $a = 0.1$, $b = 1$ and $p = 1.1$. We estimated the parameters with ten different combinations of κ and c, where κ is fixed at 0.1 and c ranges from 0.01 to 0.1. We then substituted the sets of estimates into (3.119) and plotted the resulted $Var\left[\hat{\Theta}\right]$ versus κ for a, b and p respectively in Figure 5.7. On the same plot, we show the theoretical $Var\left[\hat{\Theta}\right]$ by putting the true parameter values into (3.119) for comparison. We can see that, apart from b, the conformity between the theory and practice is satisfactory. Therefore, users can choose the set of estimates that minimise $Var\left[\hat{a}\right]$ or $Var\left[\hat{p}\right]$. Anyway, we know that the accuracy of the estimates are not very sensitive to the choice of the combination of κ and c. In general, the estimates provided by the attenuated moment estimator are consistently precise, given that the sample size is largish.

The method of attenuated moments is undoubtedly better than the MLE and the fractional moment estimator, given its outstanding performance in estimating the three parameters of a linear combination of two exponential distributions.

The Method of Appell-Fourier Moments

In Chapter 3, we presented the estimation results of a two-component positive mixture exponential distribution using the method of Appell moments with $\alpha = 3$, 4 and 5. Here, we do the same to a linear combination of two exponential distributions and follow the routine from (3.147) to (3.155) for $\alpha = 3$. We do not present the estimation results of this method with $\alpha = 4$ and $\alpha = 5$ simply because the estimators are strongly implausible for a linear combination, even when the sample size is as large as 1000. Like before, we consider three degrees of separation $r = (2, 5, 10)$ and different sample size $n_o = (10, 15, 20, 50, 1000)$. For each combination, we use 19 values of $\omega = (0.01, 0.02, ..., 0.1, 0.2, ...1)$ to estimate the three parameters of the distribution; for each ω, we calculate the bias^2, variance and mean square error of the 10000 estimates. The minimum measures of error are presented in Tables 5.22, 5.23 and 5.24 for $r = 2$, 5 and 10 respectively.

Let us first examine the performance of this method in estimating a linear distribution

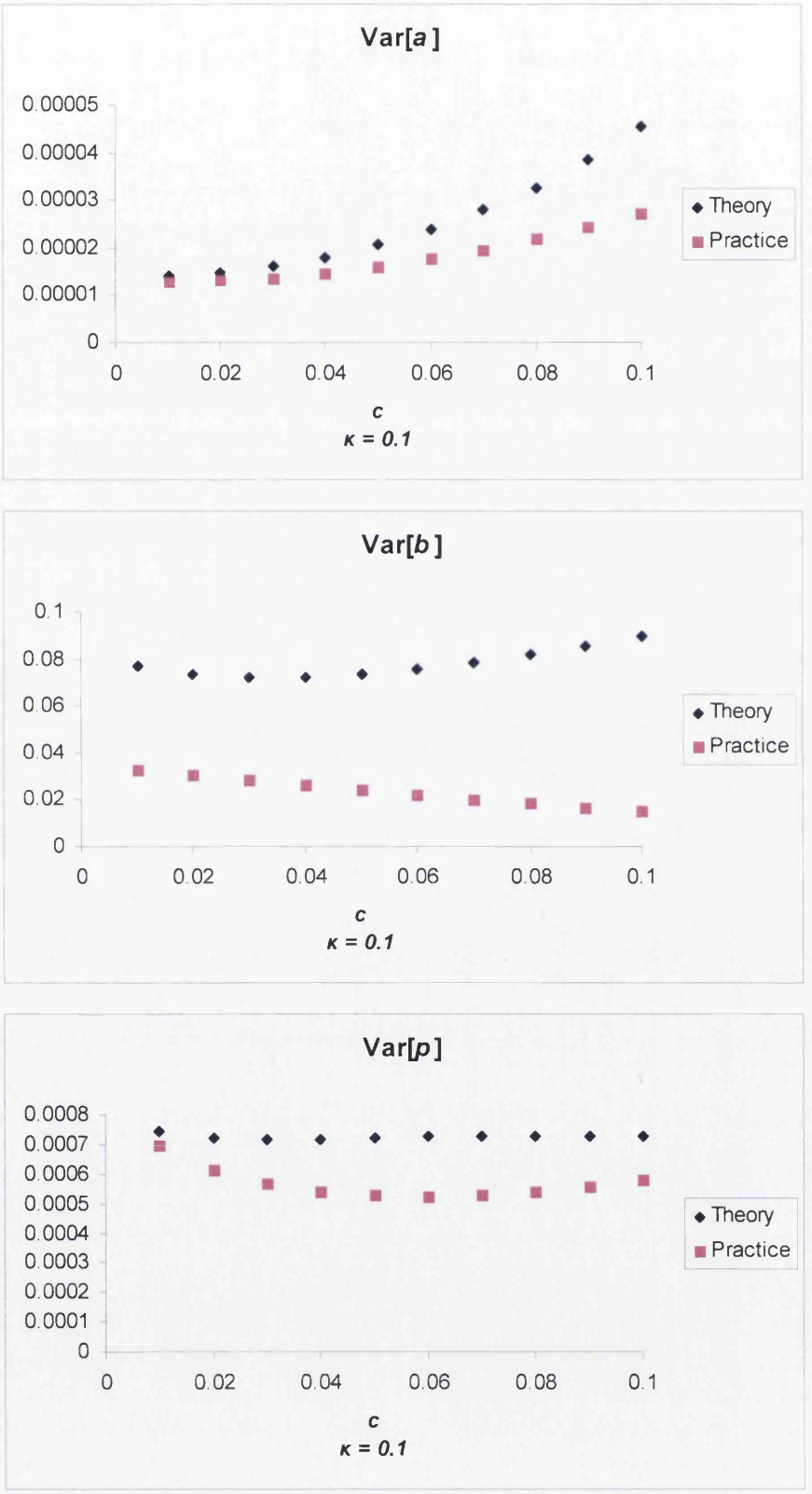


Figure 5.7: Asymptotic variance of the attenuated moment estimator given by true parameters and parameter estimates versus c, based on a data set, consisting of 1000 observations, simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 0.6$.

of two exponential distributions with $r = 2$ and $p = 1.5$ in Table 5.22. Apparently, this method is only reasonable when the number of observations in a sample is large enough. When $n_o = 1000$, any ω with a value between 0.01 and 0.08 will return reasonable estimates of Θ close to the true values. The best ω for a is 0.02 in terms of the mean square error; whereas the best ω for b is 0.06. Although the ω which minimises the variance of estimator $\hat{p}$ is 0.1, we believe that the best ω for p is 0.08 because it has the lowest bias of $\hat{p}$ while its variance of $\hat{p}$ is 1.9046, which is not much larger than the one given by $\omega = 0.1$. We do not prefer $\omega = 0.1$ essentially because it gives a highly biased $\hat{b}$ even when $n_o = 1000$. This estimator is not ideal for small samples; the estimates of b and p are either highly biased and lowly deviated, or lowly biased and highly deviated. For instance, when $n_o = 10$, the best ω for $\hat{b}$ is 0.3 in terms of bias but its variance is 4263; the best ω for $\hat{b}$ is 0.02 in terms of variance, however $E\left[\hat{b}\right]$ in this case is 0.1433 with bias^2 as 0.0032, which is a lot larger than the one given by $\omega = 0.3$. Similarly, for $\hat{p}$, when $\omega = 0.05$ is used on sample of size $n_o = 10$, $E[\hat{p}] = 1.1430$ is the lowest bias. However its variance of $\hat{p}$ is significantly large with $Var[\hat{p}] = 31$, this makes $\omega = 0.05$ implausible for the estimation of p. The variance of $\hat{p}$ is minimised when $\omega = 0.02$, however in this case $E[\hat{p}] = 0.8631$ which is not only highly biased but at the same time the estimator "sees" the distribution as a positive mixture rather than a linear combination. It appears that the best ω for estimating p, regardless of r, is either 0.01 or 0.02 for samples of small sizes. However, with these values of ω, the estimates of p are highly biased and lower than 1; for instance, when $\omega = 0.01$, $E[\hat{p}] = 0.9140$ when $n_o = 15$ whereas $E[\hat{p}] = 0.9639$ when $n_o = 20$. In other words, the fitted distribution is a positive mixture if one uses $\omega = 0.01$ or 0.02 on small samples. Therefore, we see the disagreement between the best ω in terms of bias and the best ω in terms of variance for small samples ($n_o \leq 50$) in Tables 5.22, 5.23 and 5.24.

For $r = 5$, the performance of the Appell moment estimators is similar to the ones for $r = 2$, as shown in Table 5.23. In general, to reduce the variances of the estimators, we should use small ω (≤ 0.05) to estimate the parameters. The estimation of a is satisfactory for all n_o but the estimation of b is very poor even when the sample size is as large as $n_o = 1000$. For p, again, the variance is, in general, greater than 1 for data sets with a small number of observations ($n_o \leq 50$).

Since this estimator is not good enough for small samples, we focus on its performance on data sets of large sizes with $r = 5$, $p = 1.1$ and $n_o = 1000$ in Table 5.23. The "good" ω's are 0.03, 0.04, 0.09 and 0.1 as they provide estimates with $E[\hat{a}]$, $E\left[\hat{b}\right]$ and $E[\hat{p}]$ near to the true values. Among these four ω's, the best candidate for both a and p is $\omega = 0.04$ in terms of the variance of estimator. Although $\omega = 0.03$ minimises the variance of estimator $\hat{b}$, it is significantly large ($Var\left[\hat{b}\right] = 115$) even when $n_o = 1000$.

Similarly, as seen in Table 5.24, the estimates of both b and p for $r = 10$ are considerably poor; both estimators have large variances especially when sample size is small.

When $r = 10$ and $p = 1.1$, the best ω that returns reasonable estimates of all parameters appear to have a value of 0.1. However, we note from Table 5.24 that the best ω for

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	3.07×10^{-6} (0.06)	0.00011 (0.07)	8.21×10^{-9} (0.09)	0.0002 (0.04)	1.73×10^{-7} (0.03)
$(\hat{b} - b)^2$	0.0002 (0.3)	4.92×10^{-5} (0.06)	0.0275 (0.04)	0.0002 (0.09)	0.00193 (0.01)
$(\hat{p} - p)^2$	0.1274 (0.05)	0.0579 (0.06)	0.0440 (0.09)	0.0110 (0.1)	9.81×10^{-5} (0.08)
$Var[\hat{a}]$	0.0016 (0.02)	0.0014 (0.03)	0.0012 (0.03)	0.0008 (0.03)	6.92×10^{-5} (0.02)
$Var[\hat{b}]$	48 (0.02)	147 (0.03)	474 (0.05)	18 (0.02)	1.6623 (0.06)
$Var[\hat{p}]$	1.9211 (0.02)	1.6172 (0.01)	1.4594 (0.01)	2 (0.02)	0.7564 (0.1)
$MSE[\hat{a}]$	0.0024 (0.01)	0.0019 (0.04)	0.0016 (0.04)	0.0009 (0.04)	6.96×10^{-5} (0.02)
$MSE[\hat{b}]$	48 (0.02)	147 (0.03)	474 (0.05)	18 (0.02)	1.6694 (0.06)
$MSE[\hat{p}]$	2 (0.02)	1.9605 (0.01)	1.7468 (0.01)	2.3462 (0.02)	0.7678 (0.1)

Table 5.22: Performance of the method of Appell moments (with $\alpha = 3$) for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\hat{a} - a)^2$	6.39×10^{-7} (0.06)	3.0805×10^{-5} (0.05)	4.35×10^{-6} (0.06)	1.41×10^{-8} (0.08)	1.72×10^{-7} (0.01)
$(\hat{b} - b)^2$	0.0006 (0.2)	0.1323 (0.06)	0.0168 (0.3)	0.0068 (0.2)	0.0047 (0.09)
$(\hat{p} - p)^2$	0.0022 (0.06)	0.0019 (0.09)	6.02×10^{-5} (0.06)	0.0003 (0.1)	9.53×10^{-6} (0.2)
$Var[\hat{a}]$	0.0022 (0.01)	0.0018 (0.01)	0.0016 (0.03)	0.0011 (0.05)	7.54×10^{-5} (0.04)
$Var[\hat{b}]$	114 (0.02)	38 (0.04)	35 (0.01)	27 (0.01)	115 (0.03)
$Var[\hat{p}]$	1.1727 (0.01)	1.6401 (0.02)	0.8181 (0.01)	1.2844 (0.01)	0.2645 (0.04)
$MSE[\hat{a}]$	0.0025 (0.01)	0.0022 (0.01)	0.0019 (0.04)	0.0011 (0.05)	7.77×10^{-5} (0.04)
$MSE[\hat{b}]$	114 (0.02)	39 (0.04)	35 (0.01)	27 (0.01)	115 (0.03)
$MSE[\hat{p}]$	1.2466 (0.01)	1.7255 (0.02)	0.9005 (0.01)	1.3516 (0.01)	0.2717 (0.04)

Table 5.23: Performance of the method of Appell moments (with $\alpha = 3$) for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 1.1$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	1.94×10^{-5} (0.05)	3.33×10^{-7} (0.05)	5.28×10^{-6} (0.05)	7.57×10^{-7} (0.03)	2.97×10^{-6} (0.04)
$\left(\bar{\hat{b}} - b\right)^2$	0.0066 (0.06)	0.4643 (0.06)	0.0251 (0.05)	0.0002 (0.04)	0.0006 (0.1)
$(\bar{\hat{p}} - p)^2$	3.03×10^{-6} (0.09)	0.0013 (0.05)	0.0008 (0.05)	1.10×10^{-5} (0.05)	0.0007 (0.4)
$Var[\hat{a}]$	0.0024 (0.01)	0.0019 (0.02)	0.0017 (0.03)	0.0011 (0.04)	3.83×10^{-5} (0.04)
$Var\left[\hat{b}\right]$	285 (0.01)	51 (0.04)	18 (0.03)	93 (0.01)	129 (0.02)
$Var[\hat{p}]$	1.6548 (0.02)	1.0613 (0.02)	3 (0.02)	1.4852 (0.01)	0.0372 (0.04)
$MSE[\hat{a}]$	0.0027 (0.01)	0.0022 (0.01)	0.0019 (0.04)	0.0011 (0.04)	4.13×10^{-5} (0.04)
$MSE\left[\hat{b}\right]$	286 (0.01)	52 (0.04)	19 (0.03)	94 (0.01)	129 (0.02)
$MSE[\hat{p}]$	1.7074 (0.02)	1.1174 (0.02)	3 (0.02)	1.5028 (0.01)	0.0399 (0.04)

Table 5.24: Performance of the method of Appell moments (with $\alpha = 3$) for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 1.1$ for different sample size n_o.

estimating a is 0.04, that gives estimates with both lowest bias and smallest variance. $\omega = 0.10$ returns estimates of b with the lowest bias^2, however its variance is large with a value of 1608. From the table, the best ω for $\hat{b}$, in terms of variance, is 0.02. Nevertheless, $E\left[\hat{b}\right] = 0.0526$ in this case, which is highly biased. Not to mention that the variance of $\hat{b}$ is significantly large, $Var\left[\hat{b}\right] = 129$ even when $n_o = 1000$.

To conclude, the method of Appell moments perform poorly in fitting a linear combination of two exponential distributions, especially for samples of small sizes. Even when the number of observations of a data set is as large as 1000, the estimation of b is implausible with a large variance. Therefore, we do not recommend the use of this method on the estimation problem of a linear combination of distributions.

The Method Using Order Statistics

Finally, we examine the performance of the method using order statistics, investigated in Chapter 3, in estimating the parameters of a linear combination of two exponential distributions. The estimation results are summarised in Tables 5.25, 5.26 and 5.27.

We can see that, for all r, $Var\left[\hat{b}\right]$ is extremely large, even when the sample size is large, $n_o = 1000$. This method performs badly especially when the separation between the two components is large; when $r = 10$, both the bias^2 and variance of $\hat{b}$ are large at an

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\bar{\hat{a}} - a\right)^2$	0.0003	0.0003	0.0003	0.0002	9.80×10^{-15}
$\left(\bar{\hat{b}} - b\right)^2$	0.0629	0.0048	1.1932	0.0380	0.0012
$\left(\bar{\hat{p}} - p\right)^2$	0.2499	0.2177	0.1859	0.0933	0.0457
$Var\left[\hat{a}\right]$	0.0014	0.0011	0.0009	0.0007	6.42×10^{-5}
$Var\left[\hat{b}\right]$	161	2983	5630	427	0.0302
$Var\left[\hat{p}\right]$	0.6762	0.6562	0.7319	1.6197	3
$MSE\left[\hat{a}\right]$	0.0017	0.0014	0.0012	0.0009	6.42×10^{-5}
$MSE\left[\hat{b}\right]$	162	2983	5631	427	0.0314
$MSE\left[\hat{p}\right]$	0.9261	0.8739	0.9177	1.7130	4

Table 5.25: Performance of the method using order statistics for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\bar{\hat{a}} - a\right)^2$	3.97×10^{-5}	6.73×10^{-5}	6.82×10^{-5}	3.83×10^{-5}	1.61×10^{-6}
$\left(\bar{\hat{b}} - b\right)^2$	0.7563	0.1571	0.4019	3	0.0134
$\left(\bar{\hat{p}} - p\right)^2$	0.0249	0.0139	0.0151	0.0040	0.0036
$Var\left[\hat{a}\right]$	0.0022	0.0015	0.0013	0.0008	3.85×10^{-5}
$Var\left[\hat{b}\right]$	1372	114	89	40528	534
$Var\left[\hat{p}\right]$	0.4090	1.5428	0.4526	0.3991	0.1036
$MSE\left[\hat{a}\right]$	0.0022	0.0016	0.0014	0.0008	4.02×10^{-5}
$MSE\left[\hat{b}\right]$	1373	115	90	40531	534
$MSE\left[\hat{p}\right]$	0.4339	1.5567	0.4678	0.4031	0.1072

Table 5.26: Performance of the method using order statistics for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 1.1$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	1.85×10^{-5}	2.27×10^{-5}	1.77×10^{-5}	4.83×10^{-7}	4.51×10^{-7}
$\left(\bar{\hat{b}} - b\right)^2$	43	1.0499	1.2733	5	13
$(\bar{\hat{p}} - p)^2$	0.0103	0.0043	0.0010	0.0023	0.0003
$Var[\hat{a}]$	0.0022	0.0016	0.0014	0.0007	2.31×10^{-5}
$Var\left[\hat{b}\right]$	27055	217	294	32965	96996
$Var[\hat{p}]$	0.7271	0.7513	1.2339	0.4672	0.0049
$MSE[\hat{a}]$	0.0022	0.0017	0.0014	0.0007	2.35×10^{-5}
$MSE\left[\hat{b}\right]$	270597	218	295	32970	97009
$MSE[\hat{p}]$	0.7373	0.7555	1.2349	0.4695	0.0052

Table 5.27: Performance of the method using order statistics for 10000 data sets simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 1.1$ for different sample size n_o.

unacceptable level. Because of this, the method using order statistics is outperformed by other estimators considered above.

5.1.5 Comparison of Estimation Methods

By now, it is clear that the MLE, AP and OS are implausible for the parameter estimation of a linear combination of two exponential distributions. The MLE fails to identify a linear combination in most cases. For samples of large sizes, the ML inferred distribution, regardless of the starting values $\Theta^{(0)}$, is a single distribution rather than a linear combination. In Figure 5.8, we show the scatter plots of $\hat{b}$ against $\hat{a}$ given by the MLE and the attenuated moment estimator for distributions with $r = (2, 5, 10)$ and $n_o = 1000$. It is obvious that, for each r, the scatter plots of the ML estimates show a positive linear relationship; $\hat{b}$'s have similar values to $\hat{a}$'s and never exceed 0.15 in all cases. Indeed, the scatter plot for $r = 2$ given by the MLE looks like a 45° line. Conversely, the estimates of b given by the AM are reasonable with most $\hat{b}$'s scattering in the region near to the true value. We do notice a number of cases of the under-estimation of a and over-estimation of b when $r = 5$ in the figure.

Although the Appell moment estimator and the method using order statistics do return better estimates compared to the MLE for large samples, their variances of estimator $\hat{b}$ are extremely large even when the number of observations are large enough (for instance, $n_o = 1000$). Thus these two methods are outperformed by the fractional moment estimator and the attenuated moment estimator.

Therefore, in order to find the best method for a linear combination of two exponential distributions, we only need to compare the FM and the AM in Tables 5.28 to 5.30. It is clear from these three tables that the AM outperforms the FM by returning estimates with lower bias and smaller variance of estimators regardless of the separation between the two

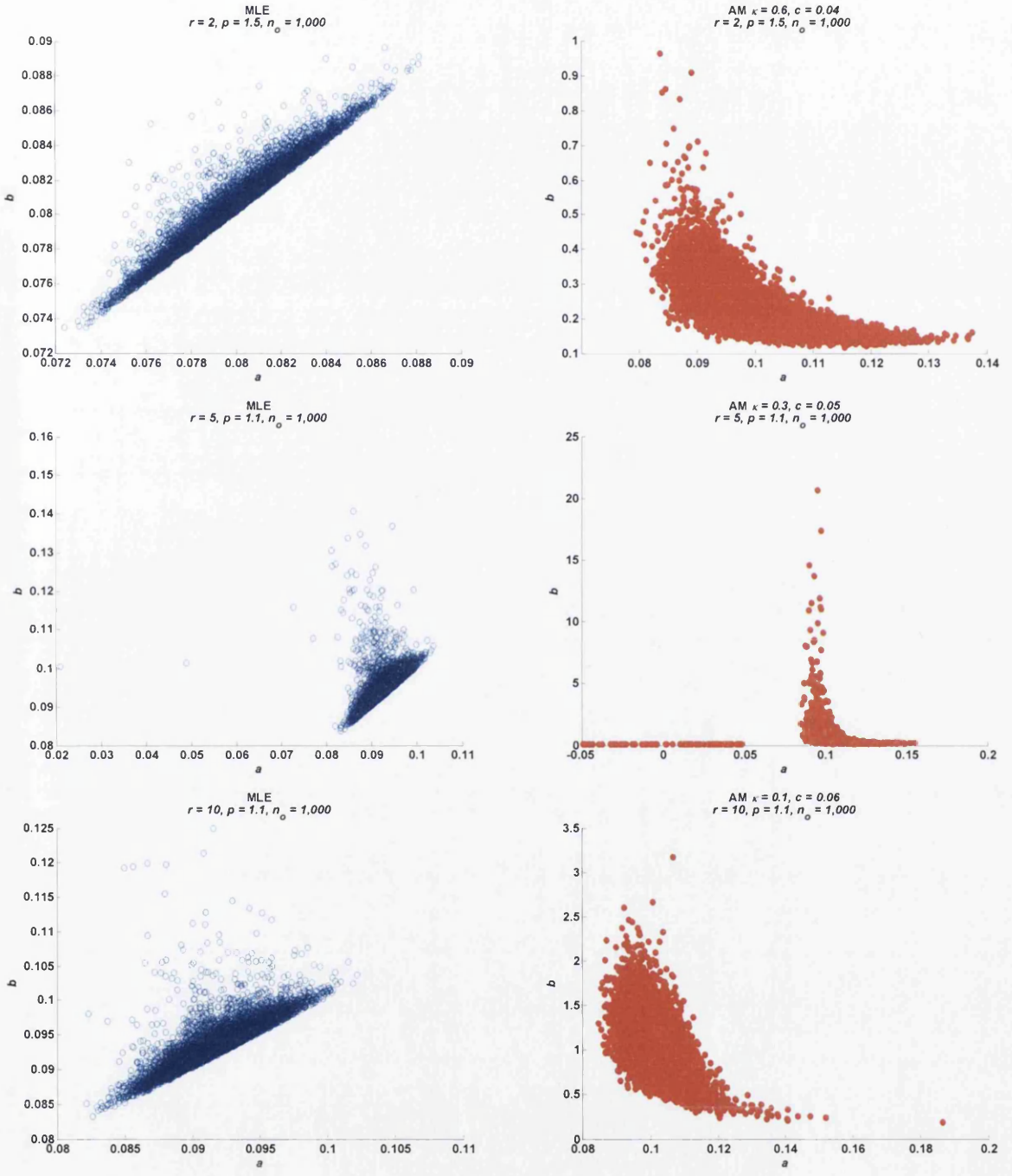


Figure 5.8: Scatter plots of b versus a: Comparison of the MLE and attenuated moment estimator for a linear combination of two exponential distributions with different r and p.

$r = 2$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
MLE	0.0004	0.0143	0.6428	4.81×10^{-6}	5.01×10^{-6}	3.19×10^{-5}
FM	4.19×10^{-10}	0.0011	0.0028	6.21×10^{-5}	0.0060	0.6564
AM	4.88×10^{-13}	0.0009	3.88×10^{-5}	6.17×10^{-5}	0.0052	0.3907
AP_3	1.73×10^{-7}	0.0019	9.81×10^{-5}	6.92×10^{-5}	1.6623	0.7564
OS	9.80×10^{-15}	0.0012	0.0457	6.42×10^{-5}	0.0302	3

Table 5.28: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.2$ and $p = 1.5$.

$r = 5$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
MLE	5.88×10^{-5}	0.1644	0.0653	8.4×10^{-6}	1.33×10^{-5}	0.0003
FM	2.51×10^{-10}	0.1136	0.0013	3.98×10^{-5}	62	0.0668
AM	2.97×10^{-11}	0.0003	1.24×10^{-5}	3.29×10^{-5}	0.4234	0.0413
AP_3	1.72×10^{-7}	0.0047	9.53×10^{-6}	7.54×10^{-5}	115	0.2645
OS	1.61×10^{-6}	0.0134	0.0036	3.85×10^{-5}	534	0.1036

Table 5.29: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 0.5$ and $p = 1.1$.

$r = 10$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
MLE	6.88×10^{-5}	0.8214	0.0265	7.12×10^{-6}	9.70×10^{-6}	4.14×10^{-5}
FM	1.79×10^{-7}	0.0033	3.04×10^{-5}	1.45×10^{-5}	0.1187	0.0011
AM	1.25×10^{-7}	0.0008	1.74×10^{-5}	1.42×10^{-5}	0.0814	0.0008
AP_3	2.97×10^{-6}	0.0006	0.0007	3.83×10^{-5}	129	0.0372
OS	4.51×10^{-7}	13	0.0003	2.31×10^{-5}	96996	0.0049

Table 5.30: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a linear combination of two exponential distributions with $a = 0.1$, $b = 1$ and $p = 1.1$.

$Eff\left[\hat{\Theta}\right]$	$r=2, p=1.5.$			$r=5, p=1.1.$			$r=10, p=1.1.$		
Method	a	b	p	a	b	p	a	b	p
FM	0.1456	0.0484	0.0002	0.2303	0.0002	0.0010	0.5862	0.0935	0.0127
AM	0.1465	0.0560	0.0003	0.2793	0.0346	0.0016	0.5986	0.1362	0.0158

Table 5.31: Efficiencies of different estimators for 10000 data sets each consisting of 1000 observations simulated from a linear combination of two exponential distributions with fixed $a = 0.1$ and varying b and p.

components.

Since the Fisher information matrix for a linear combination of two exponential distributions can be obtained explicitly with Jalali's (2008) solutions, as explained in Section 3.1.6, we shall now find the asymptotic efficiency of each estimator, denoted as $Eff\left[\hat{\Theta}\right]$, by dividing the CRLB of $Var\left[\hat{\Theta}\right]$, presented in Table 5.3, by the simulated variances of the estimators in Tables 5.28 to 5.30; the closer is $Eff\left[\hat{\Theta}\right]$ to one, the more efficient is an estimator. The efficiencies of $\hat{a}$, $\hat{b}$ and $\hat{p}$ for all three degrees of separation ($r = 2$, 5 and 10) are presented in Table 5.31. We can see that, even when the samples are as large as 1000, the efficiencies of all the estimators are low. The efficiency of $\hat{p}$ is especially low, due to the fact that the estimation is complicated by the fact that the true p is greater than one. Our suggestion is that one should not likely to clump more than two states in a level for a Markov process. Unless there is a great deal of data, for example, in the ion channel studies. By clumping states into a level for a Markov process, we should be prepared to obtain lowly efficient estimates for a linear combination of two distributions

5.1.6 Discussion

To sum up, most of the methods considered are not ideal for solving the parameter estimation problem for a linear combination of two exponential distributions. In some cases, we see poor estimation of both b and p with extremely large bias and variance of estimators. Even the MLE fails to provide good estimates for such a distribution: for data sets with large number of observations, the MLE returns $\hat{a}$ and $\hat{b}$ with a small difference between each other, and an estimate of $\hat{p}$ with a value less than 1.

Excitingly, the new method of attenuated moments stands out from the others, by providing estimators with relatively low bias and variances. In conclusion, we prefer the attenuated moment estimators to other estimators when solving the estimation problem of a linear combination of two exponential distributions, in particular because this method is the best in distinguishing a linear combination from a positive mixture of distributions.

5.2 A Linear Combination of Two Geometric Distributions

We now move on to study a linear combination of geometric distributions, the discrete analogue of the exponential case. A number of outstanding topics concerned with a linear combination of two geometric distributions will be discussed in this section. The first will be the conditions for such a distribution to satisfy. The second will be the simulation procedure for this distribution. Finally, we will study the performance of the methods investigated in Chapter 4 for estimating a linear combination of two geometric distributions, when one of the mixing weights is negative.

If N is a linear combination of N_a and N_b with weights p and q respectively, where

$$N_a \sim a\,(1-a)^{n-1} \text{ and } N_b \sim b\,(1-b)^{n-1},$$

then

$$f\,(n_i;\boldsymbol{\Theta}) = pa\,(1-a)^{n-1} + qb\,(1-b)^{n-1},\ n = 1,2,...,n,... \tag{5.13}$$

where $\boldsymbol{\Theta} = (a,b,p)$ and $p+q=1$. As discussed before, if p is less than 1, then we have the case arising from time reversible Markov models, and N has a mixture of geometric distributions. On the other hand, in a strongly irreversible case, p can be larger than 1. This means that the weight for the second geometric distribution q is negative.

5.2.1 Typology of a Linear Combination of Two Geometric Distributions

A linear combination of two geometric distributions should satisfy the following conditions:

Conditions of positivity

1. $a < b$

   As in (4.6) and since $0 < a, \bar{a} < 1$,

$$\begin{aligned} b &> a \\ &\Leftrightarrow\ a < 1 - \bar{a}^r \\ &\Leftrightarrow\ \bar{a}^r < \bar{a} \\ &\Leftrightarrow\ r > 1. \end{aligned}$$

2. $p \leq \dfrac{b}{b-a}$

   The PDF of the linear combination must be non-negative. Hence, we need to have

$$p \geq 0,$$

   and

$$pa + (1-p)\,b \geq 0.$$

This means that

$$\begin{aligned}(1-p)\,b &\geq -pa \\ &\Leftrightarrow\ 1-p \geq -\frac{pa}{b} \\ &\Leftrightarrow\ p(\frac{b-a}{b}) \leq 1.\end{aligned}$$

Hence,

$$p \leq \frac{b}{b-a}. \tag{5.14}$$

Condition of the mixture

1. $0 \leq p \leq 1$

   This is straightforward. For a mixture of two geometric distributions, p is the probability that N_i comes from the distribution N_a. A probability must be a non-negative value and could not exceed one.

Condition of linear-non-mixture

1. $1 < p \leq \frac{b}{b-a}$ (non-modal)

   Since a mixture has $p \leq 1$, for a linear combination to be a non-mixture, we need to have $p > 1$. As said before, $p \leq \frac{b}{b-a}$ in order to satisfy the condition that the PDF is always positive. Therefore, we have the above condition.

2. $\frac{b^2}{b^2-a^2} \leq p \leq \frac{b}{b-a}$ (modal)

   If n is the mode, we need to have

$$pa\bar{a}^n + (1-p)\,b\bar{b}^n \leq pa\bar{a}^{n-1} + (1-p)\,b\bar{b}^{n-1} \geq pa\bar{a}^{n-2} + (1-p)\,b\bar{b}^{n-2}.$$

   Since

$$pa\bar{a}^n + (1-p)\,b\bar{b}^n \leq pa\bar{a}^{n-1} + (1-p)\,b\bar{b}^{n-1},$$

   we know

$$pa\bar{a}^{n-1}(\bar{a}-1) \leq (1-p)\,b\bar{b}^{n-1}\left(1-\bar{b}\right)$$

$$\begin{aligned}&\Leftrightarrow\ -pa^2\bar{a}^{n-1} \leq (1-p)\,b^2\bar{b}^{n-1} \\ &\Leftrightarrow\ pa^2\bar{a}^{n-1} \geq (p-1)\,b^2\bar{b}^{n-1} \\ &\Leftrightarrow\ \frac{pa^2}{(p-1)\,b^2} \geq \left(\frac{\bar{b}}{\bar{a}}\right)^{n-1} \\ &\Leftrightarrow\ \ln\left[\frac{pa^2}{(p-1)\,b^2}\right] \geq (n-1)\ln\left[\frac{\bar{b}}{\bar{a}}\right],\end{aligned}$$

and since $\ln\left[\frac{pa^2}{(p-1)b^2}\right]$ is negative

$$\begin{aligned} n-1 &\geq \frac{\ln\left[\frac{pa^2}{(p-1)b^2}\right]}{\ln\left[\frac{\bar{b}}{\bar{a}}\right]} \\ \Leftrightarrow\ n &\geq 1+\frac{\ln\left[\frac{pa^2}{(p-1)b^2}\right]}{\ln\left[\frac{\bar{b}}{\bar{a}}\right]}. \end{aligned}$$

Similarly, from

$$pa\bar{a}^{n-1}+(1-p)\,b\bar{b}^{n-1} \geq pa\bar{a}^{n-2}+(1-p)\,b\bar{b}^{n-2},$$

we know

$$n \leq 2+\frac{\ln\left[\frac{pa^2}{(p-1)b^2}\right]}{\ln\left[\frac{\bar{b}}{\bar{a}}\right]}.$$

Therefore, n is the integer value of

$$2+\frac{\ln\left[\frac{pa^2}{(p-1)b^2}\right]}{\ln\left[\frac{\bar{b}}{\bar{a}}\right]}. \tag{5.15}$$

Note that when (5.15) is an integer, we have two consecutive modes. We will have a mode only when n is greater than 1, i.e.

$$\left[2+\frac{\ln\left[\frac{pa^2}{(p-1)b^2}\right]}{\ln\left[\frac{\bar{b}}{\bar{a}}\right]}\right] \geq 2,$$

and this means that

$$\begin{aligned} &\frac{\ln\left[\frac{pa^2}{(p-1)b^2}\right]}{\ln\left[\frac{\bar{b}}{\bar{a}}\right]} \geq 0 \\ \Leftrightarrow\ & pa^2 \leq (p-1)\,b^2 \\ \Leftrightarrow\ & p\left(b^2-a^2\right) \geq b^2 \\ \Leftrightarrow\ & p \geq \frac{b^2}{b^2-a^2}. \end{aligned}$$

r	b	$\frac{b^2}{b^2-a^2}$	$\frac{b}{b-a}$
2	0.19	1.3831	2.1111
3	0.2710	1.1576	1.5848
4	0.3439	1.0924	1.4100
5	0.4095	1.0634	1.3231
6	0.4686	1.0477	1.2713
7	0.5217	1.0381	1.2371
8	0.5695	1.0318	1.2130
9	0.6126	1.0274	1.1951
10	0.6513	1.0241	1.1814

Table 5.32: Lower and upper bounds for p in a linear combination of two geometric distributions with $a = 0.1$ and $b = 1 - 0.9^r$.

Hence

$$\frac{b^2}{b^2-a^2} \leq p \leq \frac{b}{b-a}. \tag{5.16}$$

Therefore, for a linear of two geometric distributions with a mode, p must be between $\frac{b^2}{b^2-a^2}$ and $\frac{b}{b-a}$. Table 5.32 outlines the lower bound and upper bound of p for a linear combination of two geometric distributions with a mode. Like its continuous exponential analogue, when the separation between the two components increases, the possible values of p are limited.

Figure 5.9 shows four PMF plots of a linear combination of two geometric distributions for varying r and p. When $a = 0.1$, $b = 0.19$ and $p = 1.2$ (indicated by the blue plot), the distribution has no mode because p is less than $\frac{b^2}{b^2-a^2}$; whereas the other three distributions have modes because their p is between $\frac{b^2}{b^2-a^2}$ and $\frac{b}{b-a}$.

5.2.2 Simulation of a Linear Combination of Two Geometric Distributions (Non-Mixture)

We shall discuss the simulation of a linear combination of two geometric distributions in this section. To simulate a data set arising from a distribution with PMF in the form of (5.13) where p satisfies the condition in (5.16), we generate a mixture of $N_a + N_b$ with probability (weight) π_{a+b}, and N_b with probability $1 - \pi_{a+b}$. The relation between π_{a+b} and p is as follows:

$$\pi_{a+b} = p\left(\frac{b-a}{b}\right). \tag{5.17}$$

We need to have $\pi_{a+b} \leq 1$ for it to be a probability. In order to prove this, we first find the PMF of the sum of two discrete random variables X and Y, which are both geometrically

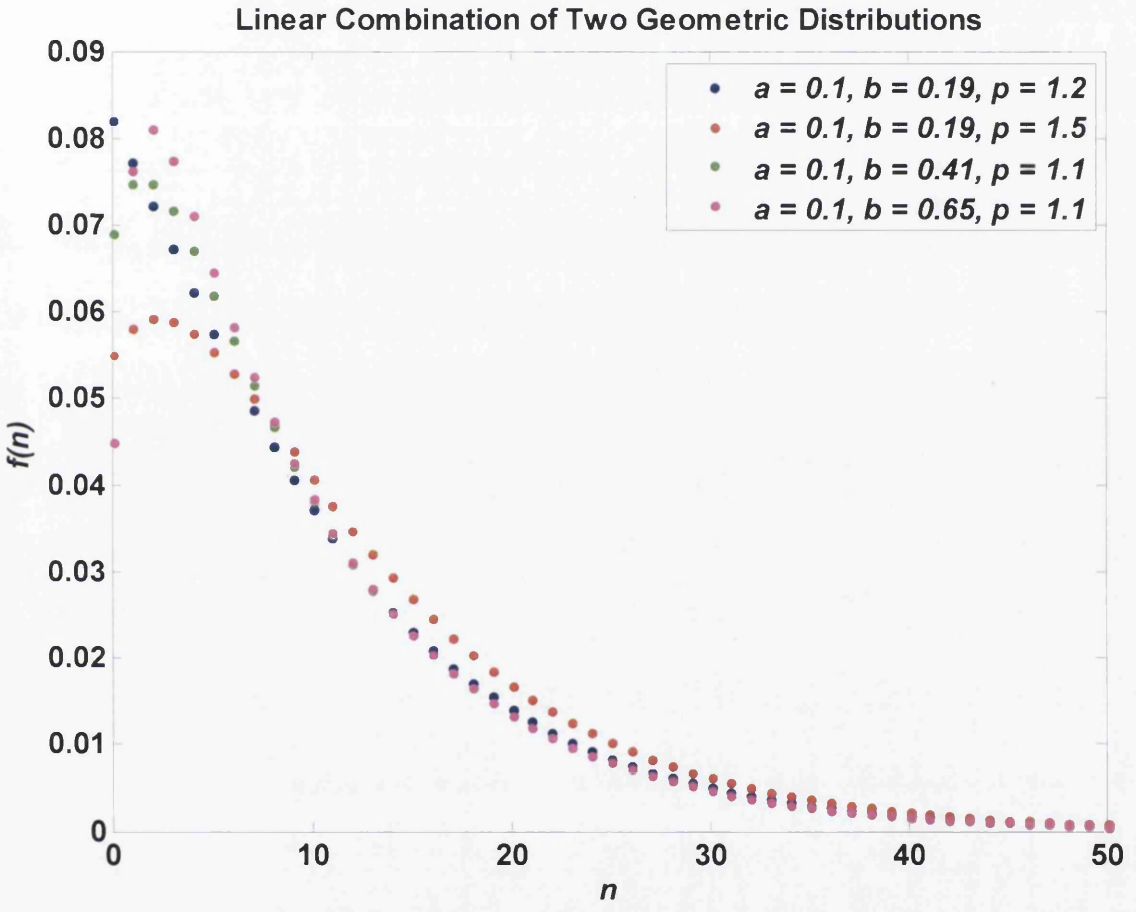


Figure 5.9: PMF plots of linear combinations of two geometric distributions for varying r and p.

distributed with distinct parameters a and b, given by

$$\begin{aligned} f_X(x) &= a(1-a)^{x-1}, \\ f_Y(y) &= b(1-b)^{y-1}, \end{aligned}$$

the PMF of the sum of these two random variables, $Z = X + Y$ is

$$f_Z(z) = \frac{ab}{b-a}\left[(1-a)^{z-1} - (1-b)^{z-1}\right].$$

Therefore, the distribution (PMF) of $N_a + N_b$ (independence of the two is assumed) is

$$\frac{ab}{b-a}(1-a)^{n-1} - \frac{ab}{b-a}(1-b)^{n-1},$$

$n = 2, 3, ...$, so the mixture of $N_a + N_b$ and N_b has the PMF

$$f(n) = \pi_{a+b}\left[\frac{ab}{b-a}(1-a)^{n-1} - \frac{ab}{b-a}(1-b)^{n-1}\right] + (1-\pi_{a+b})\,b(1-b)^{n-1}. \qquad (5.18)$$

Upon rearranging,

$$f(n) = \left[\pi_{a+b}\frac{b}{b-a}\right]a(1-a)^{n-1} + \left[1 - \pi_{a+b}\left(1 + \frac{a}{b-a}\right)\right]b(1-b)^{n-1}. \qquad (5.19)$$

It is obvious that

$$p = \pi_{a+b}\frac{b}{b-a},$$

and (5.19) is then in the form of (5.13). To check, we substitute (5.17) into (5.19),

$$\begin{aligned} f(n) &= p\left(\frac{b-a}{b}\right)\left[\frac{ab}{b-a}(1-a)^{n-1} - \frac{ab}{b-a}(1-b)^{n-1}\right] + \left(1 - p\left(\frac{b-a}{b}\right)\right)b(1-b)^{n-1} \\ &= pa(1-a)^{n-1} + \left[p\left(-\frac{a}{b} - \frac{b-a}{b}\right) + 1\right]b(1-b)^{n-1} \\ &= pa(1-a)^{n-1} + (1-p)\,b(1-b)^{n-1} \end{aligned}$$

as required.

Remark 4 *As in the case of the exponential, there is the marginal case of* $b \to a$ *and* $p \to \frac{b}{b-a}$ *and thus* $\pi_{a+b} \to 1$. *In this case the limit of the PMF* (5.18) *is the negative binomial distribution with PMF*

$$f(n) = (n-1)\,a^2(1-a)^{n-2},$$

$n = 2, 3, ...$ *Throughout this thesis, we leave aside the marginal case for our study.*

r	a	b	p
2	0.1	0.1900	1.5
5	0.1	0.4095	1.1
10	0.1	0.6513	1.1

Table 5.33: True parameters of simulated samples arising from linear combinations of two geometric distributions.

5.2.3 Estimation Methods

It is obvious that all the methods used for estimating a mixture of two geometric distributions can be applied perfectly the same way on a linear combination of two geometric distributions (since the PMF looks exactly the same). Therefore, we apply the methods discussed in Chapter 4 to simulated data sets and present the estimation results here.

Once again, simulation experiments are carried out for different sample sizes ranging from small ($n_o = 10, 15, 20, 50$) to large ($n_o = 1000$) and different separation between the components varying from small ($r = 2$) over medium ($r = 5$) to large ($r = 10$). 10000 data sets each consisting of n_o observations were simulated from a linear combination of two geometric distributions with parameter vector $\boldsymbol{\Theta} = (a, b, p)$. Table 5.33 shows the true parameters of the our simulated samples for the linear combination of two geometric distributions with different degree of separation r. The "unknown" parameters $\boldsymbol{\Theta} = (a, b, p)$ are estimated for each of the 10000 data sets based on the MLE via the EM algorithm, the method of rising factorial fractional moments (FM), the method of attenuated rising factorial fractional moments (AM) and the method of Appell Moments (APE). After the estimation procedure, we calculate the measures of error (bias^2, variance and mean square error) for all estimators to study their performances.

The Maximum Likelihood Estimator via the EM Algorithm

We simulated a data set N, with 1000 observations, arising from a linear combination of two geometric distributions with true parameters $a = 0.1$, $b = 0.6513$ and $p = 1.1$. We then fitted the data set with the MLE via the EM algorithm with initial values set at the true values $\boldsymbol{\Theta}^{(0)} = (0.1, 0.6513, 1.1)$. At each iteration k, the updated values of $\hat{a}^{(k)}$, $\hat{b}^{(k)}$ and $\hat{p}^{(k)}$ are given by (4.11), (4.12) and (4.10) respectively. The EM iteration process stopped after 42 iterations with the final estimates $\hat{\boldsymbol{\Theta}}^{(42)} = (0.0925, 0.0939, 0.6770)$ and the log-likelihood is maximised at $\hat{l}^{(42)} = -3327$. Figure 5.10 shows the value of the estimates $\hat{a}^{(k)}$, $\hat{b}^{(k)}$ and $\hat{p}^{(k)}$ at each iteration k, whereas Table 5.34 shows the first seven iterative values of the parameters and the log-likelihood function. We can see that all estimates increased gradually at the beginning but dropped dramatically at the fifth iteration. Like the exponential case, the reason behind this is the log-likelihood became a complex number at the third iteration (see Table 5.34) because the estimate of the mixing weight $\hat{p}^{(3)}$ exceeded the upper bound of p ($\frac{\hat{b}^{(3)}}{\hat{b}^{(3)} - \hat{a}^{(3)}}$) and hence the condition of positivity was violated at this point. *As a result, the*

k	$\hat{a}^{(k)}$	$\hat{b}^{(k)}$	$\hat{p}^{(k)}$	$\hat{l}^{(k)}$
0	0.1000	0.6513	1.1000	-3312
1	0.1016	0.6716	1.1088	-3313
2	0.1033	0.7146	1.1295	-3324
3	0.1118	0.8279	1.2336	$-3212+163i$
4	0.0799	1.3450	0.8502	$-3565+217i$
5	0.0285	0.5141	0.2661	-4111
6	0.0638	0.2526	0.5798	-3400
7	0.0717	0.1918	0.6345	-3354

Table 5.34: The iterative values when estimating a sample arising from a linear combination of two geometric distributions using the maximum likelihood estimator via the EM algorithm $a = 0.1$,b=0.6513,p=1.1andn$_o = 1000$. Starting values are set as true values.

log-likelihood does not converge monotonically but decreases, as shown in the table, after the violation. We can see from Figure 5.10 that after the fifth iteration, $\hat{a}^{(k)}$ and $\hat{b}^{(k)}$ converged to the same point while $\hat{p}^{(k)}$ settled at 0.6770 when the iteration process terminated. Since the estimates of a and b are similar, the ML fitted distribution of the data set is reduced to a single geometric distribution with $\hat{\theta} \approx 0.09$.

We changed the initial values to $\mathbf{\Theta}^{(0)} = (0.1, 0.6513, 0.6)$ where $p^{(0)}$ is less than 1 and hence the mixing weight of the second component is not a negative number, then we restarted the ML estimation of the sample. In Figure 5.11, we see the iterative values of the parameters at each iteration, whereas in Figure 5.12 we show the iterative value of the log-likelihood at each iteration. After the first iteration, $\hat{b}^{(k)}$ decreased and stopped at a similar point to $\hat{a}^{(k)}$. We also note that $\hat{p}^{(k)}$ increased gradually and terminated at $\hat{p}^{(39)} = 0.9321$; in Figure 5.12, we see that the log-likelihood function increased monotonically and stopped at $\hat{l}^{(k)} = -3327$. With a different set of starting values, the ML fitted distribution remains as a single geometric distribution.

We study further the behaviour of the MLE on a linear combination of two geometric distributions by carrying out simulation experiments for different sample sizes $n_o = (10, 15, 20, 50, 1000)$ and different degrees of separation $r = (2, 5, 10)$. For each r and n_o, we simulated 10000 data sets and estimated every data set with the MLE. The starting points of a and b are set at the true values but we used $p^{(0)} = 0.6$ for each sample to avoid the log-likelihood function becoming a complex number. The results are shown in Tables 5.35, 5.36 and 5.37.

It is obvious from these three tables, that the average of estimates of b decreases when sample size increases, whereas the average of estimates $\hat{p}$ increases, but never exceeds 1, for increasing sample size. For small samples, the fitted distribution is always a positive mixture. When $n_o = 1000$, $\hat{b}$ is very close to $\hat{a}$, regardless of the separation between the two components, and hence is highly biased, especially for large r, with an extremely low variance. This suggests that a majority of the 10000 estimates of b are close to $\hat{a}$ and hence the fitted distribution is a single geometric distribution rather than a linear combination.

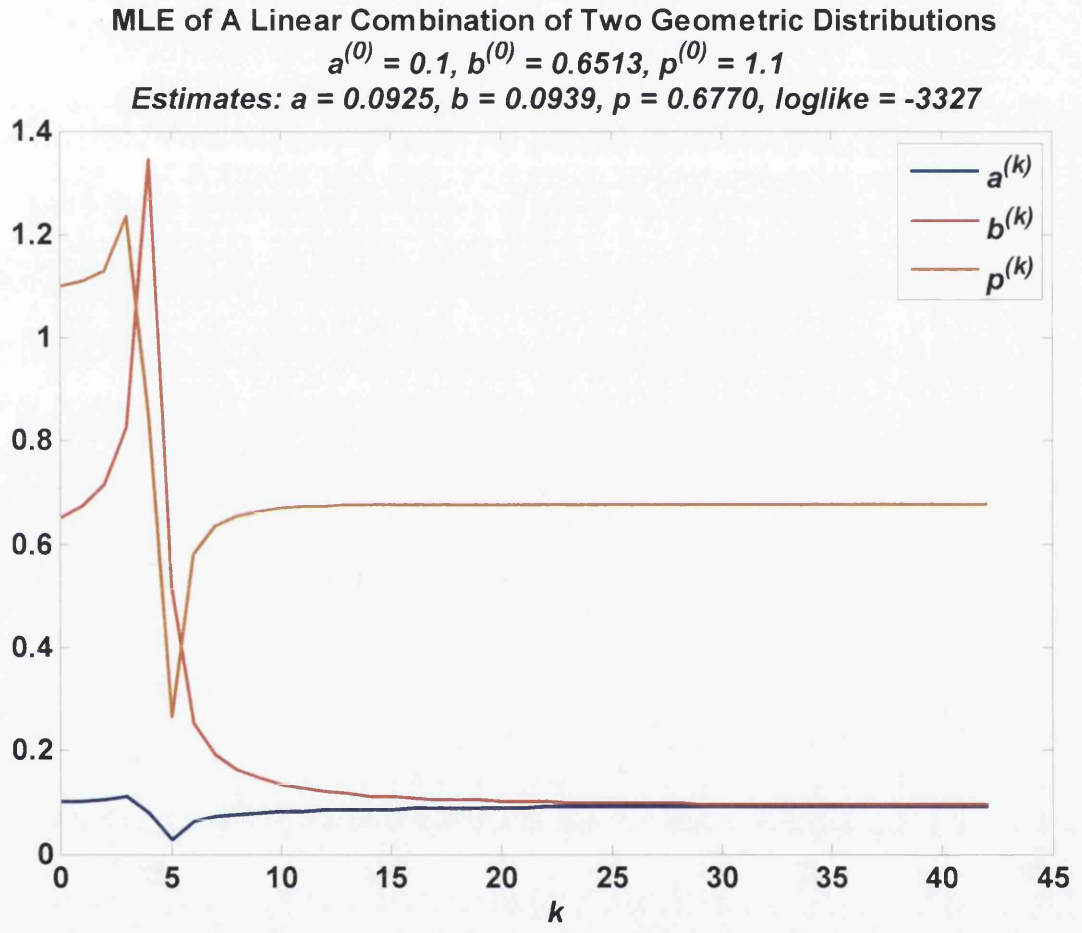


Figure 5.10: The ML updated estimates $\hat{\Theta}^{(k)}$ at each iteration k for an artificial data set, consisting 1000 observations, simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$. Starting values are set as true values.

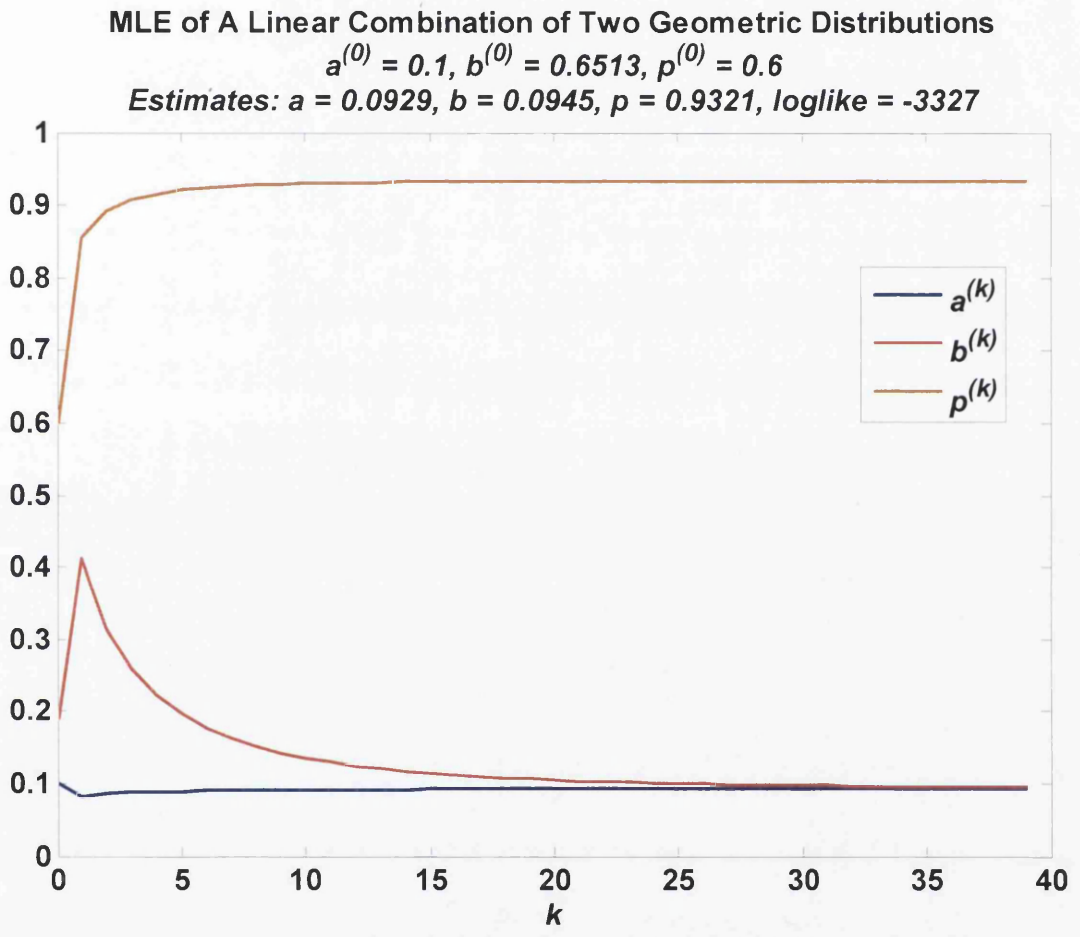


Figure 5.11: The ML updated estimates $\hat{\Theta}^{(k)}$ at each iteration k for an artificial data set, consisting 1000 observations, simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$. Starting values are $a^{(0)} = 0.1$, $b^{(0)} = 0.6513$ and $p^{(0)} = 0.6$.

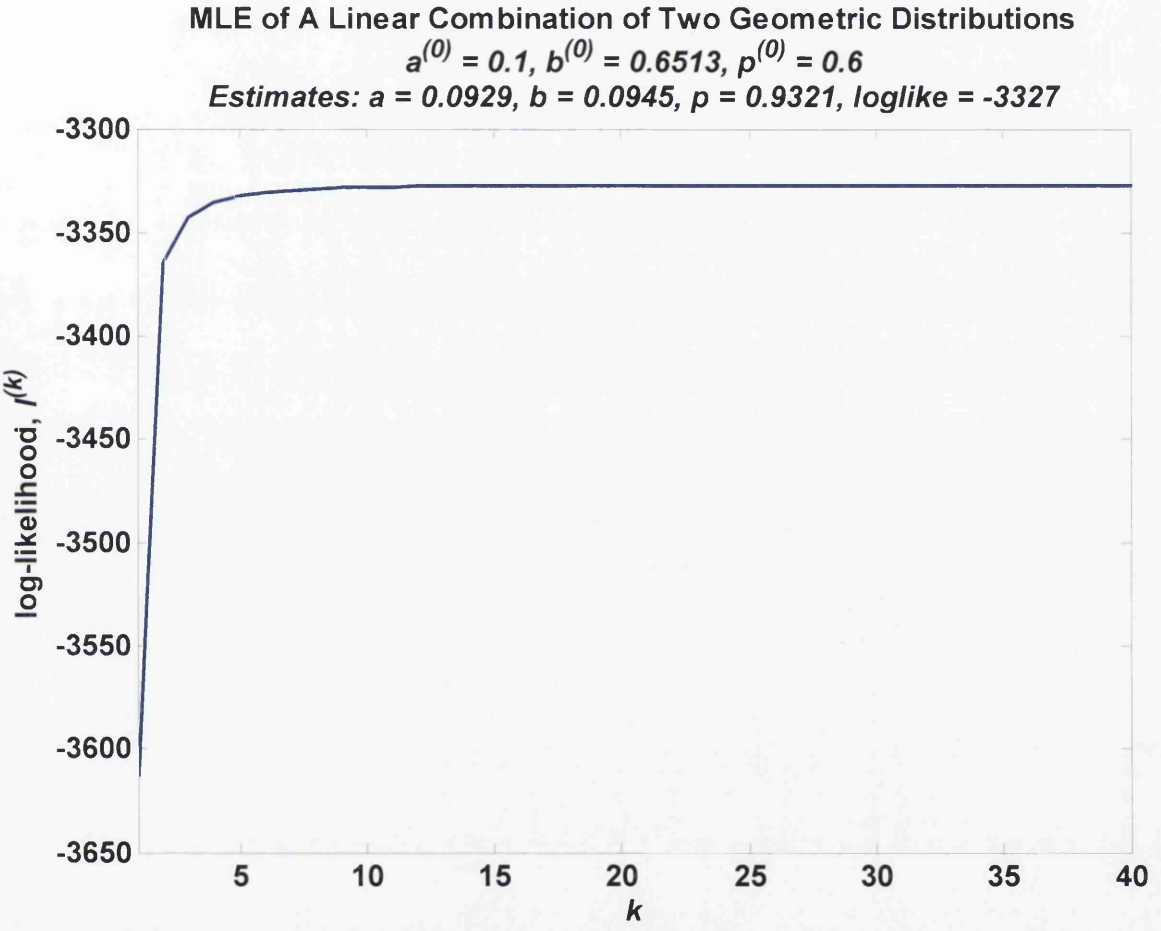


Figure 5.12: The ML updated estimate of log-likelihood $\hat{l}^{(k)}$ at each iteration k for a data set, consisting of 1000 observations, simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$. Starting values are $a^{(0)} = 0.1$, $b^{(0)} = 0.6513$ and $p^{(0)} = 0.6$.

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.0826	0.0813	0.0805	0.0800	0.0807
$E[\hat{b}]$	0.1280	0.1163	0.1082	0.0905	0.0814
$E[\hat{p}]$	0.6876	0.6879	0.6882	0.6884	0.6982
$(\hat{a} - a)^2$	0.0003	0.0004	0.0004	0.0004	0.0004
$(\hat{b} - b)^2$	0.0038	0.0054	0.0067	0.0099	0.0118
$(\hat{p} - p)^2$	0.6599	0.6596	0.6590	0.6586	0.6430
$Var[\hat{a}]$	0.0007	0.0004	0.0003	0.0001	4.56×10^{-6}
$Var[\hat{b}]$	0.0202	0.0146	0.0109	0.0029	4.79×10^{-6}
$Var[\hat{p}]$	0.0103	0.0099	0.0095	0.0076	3.25×10^{-5}
$MSE[\hat{a}]$	0.0010	0.0008	0.0007	0.0005	0.0004
$MSE[\hat{b}]$	0.0241	0.0200	0.0176	0.0128	0.0118
$MSE[\hat{p}]$	0.6702	0.6695	0.6684	0.6662	0.6430

Table 5.35: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 1.5$ for different sample size n_o. Starting values set as true values.

With these behaviours in mind, we conclude that the MLE is not an ideal method for the estimation problem of a linear combination of two geometric distributions.

The Method of Rising Factorial Fractional Moments

Let us now study the performance of the method of rising factorial fractional moments on a linear combination of two geometric distributions. We apply this method to the estimation problem by following the procedures outlined in Section 4.3.1 and considering ten fractions $\kappa = (0.1, 0.2, ..., 1)$. For each κ, we generated 10000 simulated data sets each consisting of n_o observations from a linear combination of two geometric distributions. As before, we consider three sets of parameters $\boldsymbol{\Theta} = (0.1, 0.19, 1.5)$, $\boldsymbol{\Theta} = (0.1, 0.4095, 1.1)$ and $\boldsymbol{\Theta} = (0.1, 0.6513, 1.1)$, representing different levels of separation between the components, and five sample size $n_o = (10, 15, 20, 50, 1000)$. Our objective is to find the best values of the fractions that provide estimates with high precision for different degree of separation between the two populations. For each 10000 estimates, we calculated the measures of errors and the minimum values are presented in Tables 5.38 to 5.40, where the counterpart κ are shown in brackets below the measures of errors. We shall now study the characteristics of this estimator from these tables.

Table 5.38 presents the measures of errors for linear combination of two geometric distributions with $r = 2$, $p = 1.5$ and different sample size n_o. As seen in the table, the best κ for estimating a and p, in terms of the mean square error, has a value between 0.1 to 0.5. On the other hand, the best κ for b is either 0.9 or 1 when the sample size is small, whereas the best κ for b is 0.3 when $n_o = 1000$. At the first glance, we might think that we should

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.0930	0.0912	0.0906	0.0897	0.0927
$E[\hat{b}]$	0.2281	0.2077	0.1916	0.1424	0.0948
$E[\hat{p}]$	0.8009	0.8061	0.8022	0.8083	0.8442
$(\hat{a}-a)^2$	4.85×10^{-5}	7.76×10^{-5}	8.82×10^{-5}	0.0001	5.30×10^{-5}
$(\hat{b}-b)^2$	0.0329	0.0408	0.0475	0.0713	0.0990
$(\hat{p}-p)^2$	0.0895	0.0864	0.0887	0.0851	0.0654
$Var[\hat{a}]$	0.0010	0.0007	0.0005	0.0003	7.16×10^{-6}
$Var[\hat{b}]$	0.0632	0.0561	0.0466	0.0216	1.09×10^{-5}
$Var[\hat{p}]$	0.0250	0.0251	0.0276	0.0257	0.0003
$MSE[\hat{a}]$	0.0011	0.0008	0.0006	0.0004	6.02×10^{-5}
$MSE[\hat{b}]$	0.0961	0.0969	0.0941	0.0929	0.0990
$MSE[\hat{p}]$	0.1145	0.1115	0.1162	0.1108	0.0657

Table 5.36: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 1.1$ for different sample size n_o. Starting values $\boldsymbol{\Theta}^{(0)} = (0.1, 0.4095, 0.6)$

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.0938	0.0918	0.0912	0.0906	0.0922
$E[\hat{b}]$	0.2461	0.2145	0.1888	0.1302	0.0941
$E[\hat{p}]$	0.8811	0.8835	0.8829	0.8987	0.9374
$(\hat{a}-a)^2$	3.81×10^{-5}	6.76×10^{-5}	7.84×10^{-5}	8.80×10^{-5}	6.14×10^{-5}
$(\hat{b}-b)^2$	0.1642	0.1908	0.2139	0.2716	0.3105
$(\hat{p}-p)^2$	0.0479	0.0469	0.0472	0.0405	0.0265
$Var[\hat{a}]$	0.0010	0.0007	0.0005	0.0002	6.44×10^{-6}
$Var[\hat{b}]$	0.0755	0.0612	0.0480	0.0165	9.00×10^{-6}
$Var[\hat{p}]$	0.0268	0.0290	0.0307	0.0247	3.73×10^{-5}
$MSE[\hat{a}]$	0.0010	0.0007	0.0006	0.0003	6.78×10^{-5}
$MSE[\hat{b}]$	0.2397	0.2520	0.2619	0.2880	0.3105
$MSE[\hat{p}]$	0.0747	0.0759	0.0779	0.0652	0.0265

Table 5.37: Performance of the MLE via the EM algorithm for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$ for different sample size n_o. Starting values $\boldsymbol{\Theta}^{(0)} = (0.1, 0.19, 0.6)$

$r=2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}}-a)^2$	0.0004 (0.2)	0.0004 (0.1)	0.0004 (0.2)	0.0002 (0.5)	7.55×10^{-10} (0.9)
$\left(\bar{\hat{b}}-b\right)^2$	0.0122 (1)	0.0015 (1)	0.0024 (1)	0.0545 (1)	0.0003 (1)
$(\bar{\hat{p}}-p)^2$	0.1639 (0.3)	0.1225 (0.3)	0.1514 (0.6)	0.0396 (0.6)	0.0143 (0.1)
$Var[\hat{a}]$	0.0017 (1)	0.0014 (1)	0.0013 (1)	0.0010 (0.5)	5.46×10^{-5} (0.1)
$Var\left[\hat{b}\right]$	271 (1)	77 (0.9)	149 (1)	171 (0.9)	0.0034 (0.2)
$Var[\hat{p}]$	1.3074 (0.2)	1.1300 (0.2)	1.2973 (0.5)	0.7603 (0.2)	1.1399 (0.1)
$MSE[\hat{a}]$	0.0023 (0.2)	0.0019 (0.1)	0.0019 (0.2)	0.0012 (0.5)	5.57×10^{-5} (0.1)
$MSE\left[\hat{b}\right]$	271 (1)	77 (0.9)	149 (1)	172 (0.9)	0.0040 (0.3)
$MSE[\hat{p}]$	1.5184 (0.2)	1.3142 (0.2)	1.4766 (0.5)	0.8801 (0.2)	1.1542 (0.1)

Table 5.38: Performance of the method of rising factorial fractional moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a=0.1$, $b=0.19$ and $p=1.5$ for different sample size n_o.

use a large fraction ($\kappa=0.9,1$) for small samples because they have the lowest variance of estimator $\hat{b}$. However, we found that most of the 10000 estimates of b are negative when $\kappa=1$ and this is the reason why it has both the lowest bias^2 and variance. Given that the true b is not negative, $\kappa=1$ is not the best fraction for estimating b. Nevertheless, we should note that negative b is possible in data coming from clumped Markov chains, as in the case a and b are eigenvalues of a sub-stochastic 2×2 matrix. The largest eigenvalue is positive and less than one, but the smaller eigenvalue need only to have an absolute value less than the value of the largest eigenvalue.

In fact, a low fraction $\kappa\leq0.5$ is ideal for the estimation of b; the reason they have large variance of $\hat{b}$ is that there are a few extremely large estimates of b in the 10000 estimates. As explained before, b is given by $y^{-\frac{1}{\kappa}}$ and is very sensitive to y; a small departure of $\hat{y}$ from the true value can make $\hat{b}$ extremely large. In Figure 5.13, we see a plot of $\hat{b}$ versus $\hat{y}$ when $\kappa=0.3$ is used to estimate 10000 data sets, each with 10 observations, arising from a linear combination of two geometric distributions with true parameters $a=0.1$, $b=0.19$ and $p=1.5$; the true value of y is 1.6458. As seen from the plot, when $\hat{y}$ is near to zero, $\hat{b}$ is extremely large. Since the number of observations in the data set is scarce, the probability of obtaining a poor estimate of y is greater. This is why the variance of $\hat{b}$ is extremely large when fractional moments are used on small samples. When the number of observations is as large as 1000, $Var\left[\hat{b}\right]$ is greatly reduced to 0.0034. The histogram of $\hat{b}$ is shown in Figure

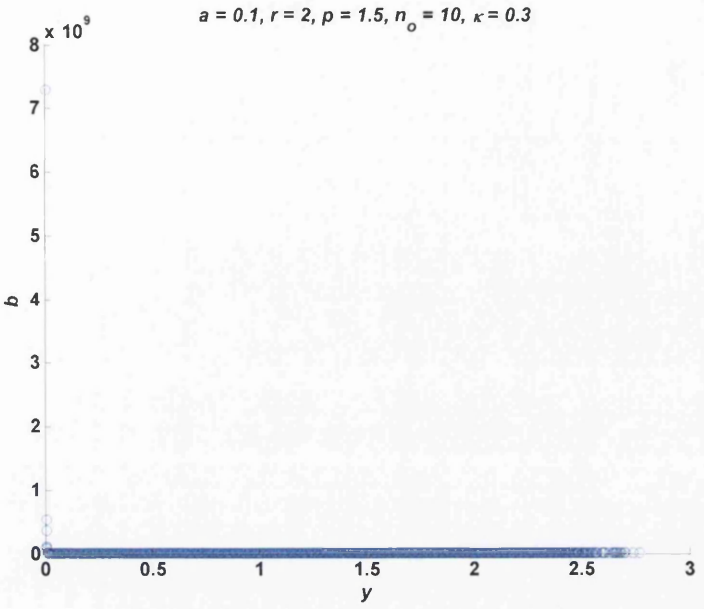


Figure 5.13: Plot of $\hat{b}$ versus $\hat{y}$ when $\kappa = 0.3$ is used to estimate 10000 data sets, each consisting of 10 observations, arising from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 1.5$.

5.14, we can see that the $\hat{b}$ is between 0.0675 and 0.8363, and most of $\hat{b}$ are close to the true values 0.19.

In general, most of the estimates of b given by any κ below 0.5 are close to the true value when $n_o = 1000$ and hence $\left(\overline{\hat{b}} - b\right)^2$ are considerably low with these low fractions. Figure 5.15 shows the box plots of $\hat{b}$ given by five different κ. It is obvious that $E\left[\hat{b}\right]$ is close to the true value in each case. However, the variance of $\hat{b}$ increases with κ. We can see from Table 5.38 that the best fraction for b, in terms of variance, is 0.2, whereas the best fraction for b is 0.3 in terms of mean square error. Therefore, we should use small κ to estimate b for such a distribution.

Table 5.39 shows the estimation results for $r = 5$ and $p = 1.1$. It is clear from the table that, for all three parameters, the best κ for large samples with $n_o = 1000$ is either 0.1 or 0.2. Our simulation experiment shows that a large fraction ($0.5 \leq \kappa \leq 1$) gives small variances when sample size is small. However, we investigated and found that these κ's actually return quite unreasonable estimates of the parameters; for instance, when $\kappa = 1$, we obtained a large number of $\hat{b}$ with negative values. Also, large κ tends to recognise most of the distribution as a positive mixture, rather than allowing negative weight. We can see from the table that $\kappa \leq 0.5$ gives a minimum $\left(\overline{\hat{p}} - p\right)^2$, $Var\left[\hat{p}\right]$ and $MSE\left[\hat{p}\right]$. Nevertheless, when small κ is used, in some data sets we obtained extremely large estimates of b, due to small $\hat{y}$, and this makes $Var\left[\hat{b}\right]$ exceptionally large.

Similarly, it is clear from Table 5.40 that the optimal κ for $r = 10$ and $p = 1.1$ is 0.1

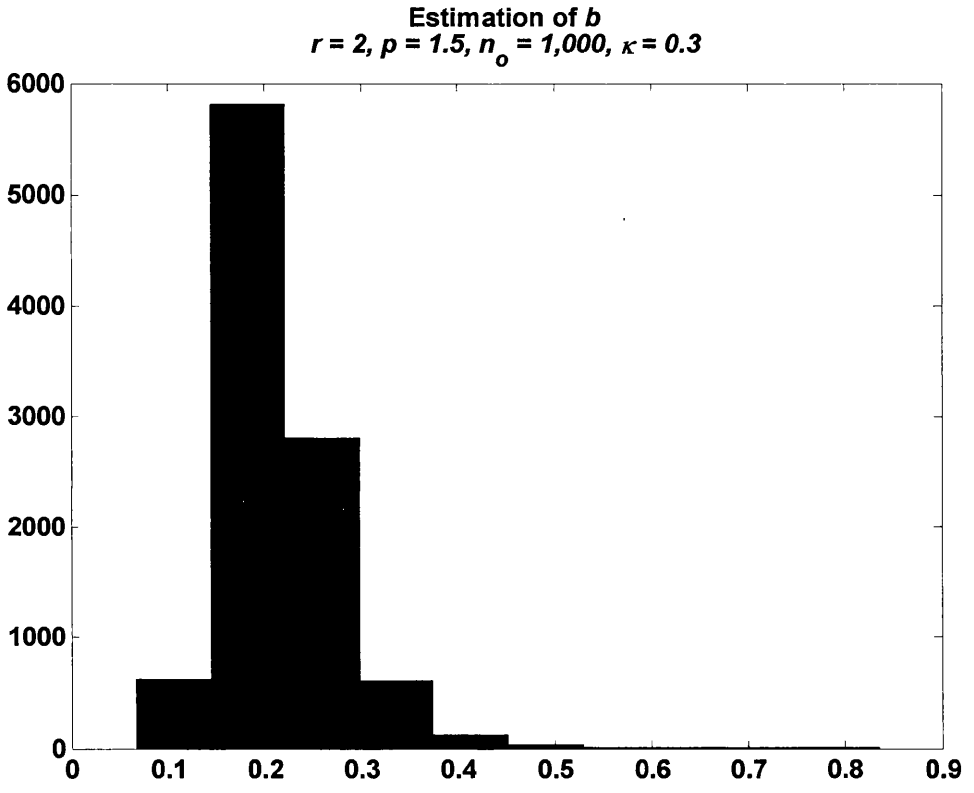


Figure 5.14: Histogram of rising factorial fractional moment estimator $\hat{b}$ when $\kappa = 0.3$ is used to estimate 10000 data sets, each consisting of 1000 observations, simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 1.5$.

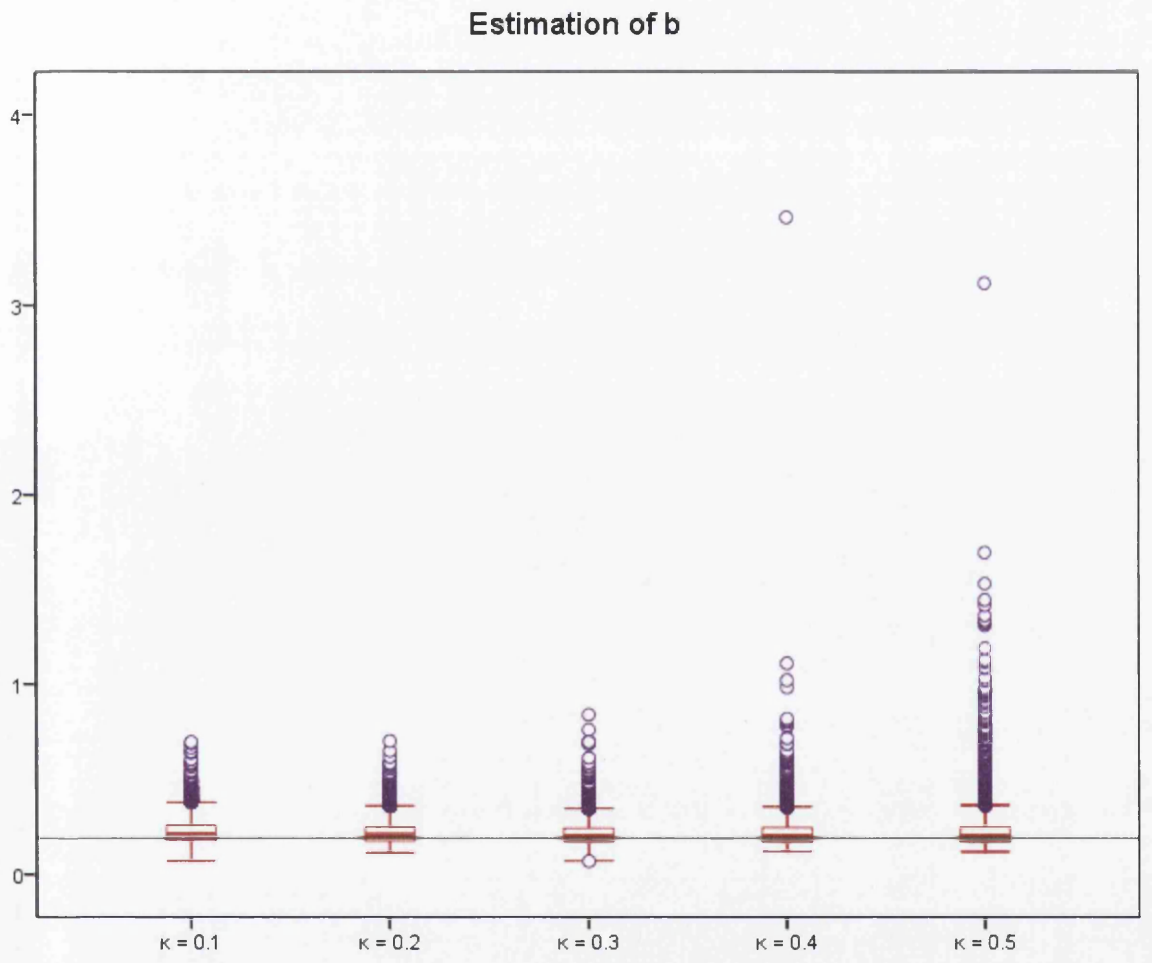


Figure 5.15: Box plots of $\hat{b}$ for various κ used to estimate 10000 data sets, each consisting of 1000 observations, arising from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$, $p = 1.5$.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	8.05×10^{-5} (0.1)	0.0001 (0.2)	0.0001 (0.1)	9.66×10^{-5} (0.1)	1.43×10^{-6} (0.2)
$\left(\bar{\hat{b}} - b\right)^2$	0.0745 (1)	0.1439 (1)	0.1887 (1)	0.0996 (0.9)	0.0108 (0.2)
$(\bar{\hat{p}} - p)^2$	0.0056 (0.2)	0.0066 (0.2)	0.0075 (0.1)	0.0032 (0.4)	0.0023 (0.1)
$Var[\hat{a}]$	0.0020 (1)	0.0017 (1)	0.0015 (1)	0.0012 (1)	2.98×10^{-5} (0.2)
$Var\left[\hat{b}\right]$	137 (1)	153 (1)	268 (1)	125 (0.9)	0.3756 (0.2)
$Var[\hat{p}]$	1.1530 (0.5)	0.8889 (0.5)	0.7362 (0.1)	0.5696 (0.2)	0.0499 (0.1)
$MSE[\hat{a}]$	0.0024 (1)	0.0020 (1)	0.0019 (0.1)	0.0014 (0.2)	3.13×10^{-5} (0.2)
$MSE\left[\hat{b}\right]$	137 (1)	153 (1)	269 (1)	125 (0.9)	0.3864 (0.2)
$MSE[\hat{p}]$	1.1744 (0.5)	0.9163 (0.1)	0.7437 (0.1)	0.5751 (0.2)	0.0522 (0.1)

Table 5.39: Performance of the method of rising factorial fractional moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 1.1$ for different sample size n_o.

for all three parameters when $n_o = 1000$. The mean square errors of a and p are minimised when κ is either 0.1 or 0.2 regardless of the sample size n_o. Although the variance of $\hat{b}$ is minimised when $\kappa = 0.9$ or 1 for samples of small sizes, the fitted distribution is a positive mixture rather than a linear combination in most cases. The parameter estimates given by small fraction are more reasonable. However, the main drawback is that we might get an extremely large estimate of b if $\hat{y}$ is deviated from the true value.

The method of rising factorial fractional moments is a good method for the estimation problem of a linear combination of two geometric distributions, provided that the number of observations in a data set is large enough. Tables 5.41, 5.42 and 5.43 compare the practical and theoretical minimum variances of the estimators (given by (4.27)) and their counterpart fraction κ. We used the theoretical optimal κ to estimate another 10000 data sets simulated from the specified distribution and recorded the observed variances in brackets underneath the theoretical values. As seen in these tables, the approximation given by (4.27) underestimates the practical $Var\left[\hat{\boldsymbol{\Theta}}\right]$. As said before, (4.27) is only an approximation to the theoretical values of $Var\left[\hat{\boldsymbol{\Theta}}\right]$, there exists approximation errors in the computation and random errors in the simulation experiments. The fact that q is negative has caused more complication to the approximation and hence we do not find good conformity between theory and practice.

However, the best κ suggested by the theory does return estimates with low variances;

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	1.19×10^{-5} (0.1)	1.05×10^{-5} (0.1)	7.53×10^{-6} (0.1)	3.40×10^{-7} (0.3)	2.81×10^{-7} (0.2)
$\left(\bar{\hat{b}} - b\right)^2$	0.3126 (1)	0.0031 (0.9)	0.0008 (0.9)	0.0607 (0.9)	0.0012 (0.1)
$(\bar{\hat{p}} - p)^2$	1.84×10^{-7} (0.1)	3.88×10^{-5} (0.1)	9.43×10^{-6} (0.6)	1.45×10^{-5} (0.8)	5.91×10^{-5} (0.1)
$Var[\hat{a}]$	0.0021 (1)	0.0018 (1)	0.0016 (1)	0.0008 (0.1)	1.49×10^{-5} (0.1)
$Var\left[\hat{b}\right]$	481 (1)	56 (0.9)	44 (0.9)	19 (1)	0.0490 (0.1)
$Var[\hat{p}]$	0.9553 (0.1)	0.6615 (0.1)	1.1187 (0.2)	0.3123 (0.1)	0.0016 (0.1)
$MSE[\hat{a}]$	0.0024 (1)	0.0019 (0.1)	0.0017 (0.1)	0.0008 (0.1)	1.52×10^{-5} (0.1)
$MSE\left[\hat{b}\right]$	482 (1)	56 (0.9)	44 (0.9)	19 (1)	0.0502 (0.1)
$MSE[\hat{p}]$	0.9553 (0.1)	0.6615 (0.1)	1.1212 (0.2)	0.3165 (0.1)	0.0016 (0.1)

Table 5.40: Performance of the method of rising factorial fractional moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$ for different sample size n_o.

r	p	Theoretical Optimal κ	Theoretical $Var[\hat{a}]$	Practical Optimal κ	Practical $Var[\hat{a}]$
2	1.5	0.3480	8.83×10^{-5} (5.66×10^{-5})	0.1	5.46×10^{-5}
5	1.1	0.0572	2.02×10^{-5} (3.70×10^{-5})	0.2	2.98×10^{-5}
10	1.1	−0.0916	1.29×10^{-5} (1.41×10^{-5})	0.1	1.49×10^{-5}

Table 5.41: Theoretical and simulated minimum variance of the rising factorial fractional moment estimator $\hat{a}$ given by the optimal κ for a linear combination of two geometric distributions with fixed $a = 0.1$ and $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

r	p	Theoretical Optimal κ	Theoretical $Var\left[\hat{b}\right]$	Practical Optimal κ	Practical $Var\left[\hat{b}\right]$
2	1.5	0.3553	0.0036 (0.0044)	0.2	0.0034
5	1.1	0.0535	0.0303 (0.0786)	0.2	0.3756
10	1.1	−0.1930	0.0196 (0.0215)	0.1	0.0490

Table 5.42: Theoretical and simulated minimum variance of the rising factorial fractional moment estimator $\hat{b}$ given by the optimal κ for a linear combination of two geometric distributions with fixed $a = 0.1$ and $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

r	p	Theoretical Optimal κ	Theoretical $Var\left[\hat{p}\right]$	Practical Optimal κ	Practical $Var\left[\hat{p}\right]$
2	1.5	0.3510	0.2053 (2.354)	0.1	1.1399
5	1.1	0.0375	0.0038 (0.1483)	0.1	0.0499
10	1.1	−0.1773	0.0009 (0.0012)	0.1	0.0016

Table 5.43: Theoretical and simulated minimum variance of the rising factorial fractional moment estimator $\hat{p}$ given by the optimal κ for a linear combination of two geometric distributions with fixed $a = 0.1$ and $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

For $r = 5$ and 10, the $Var\left[\hat{b}\right]$ given by the κ's suggested (both lower than 0.1) are in fact lower than the ones we obtained before. We did not use negative fraction in our simulation experiments but, surprisingly, the negative κ suggested by the theory for $r = 10$ does in fact provide estimates of $\boldsymbol{\Theta}$ with lower variances that the minimum variances we obtained from our previous simulation experiments. We note from these tables that, the optimal κ is always small (≤ 0.2) and decreases with r.

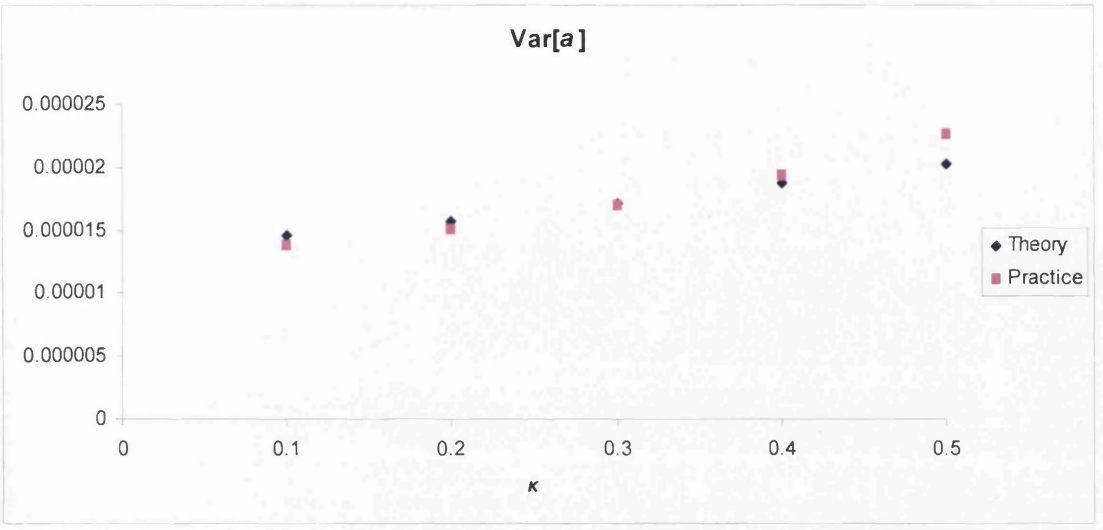
In Chapter 4, we showed how one should choose an appropriate κ when the degree of separation between the components is not known in practice. We shall now demonstrate that the approach can also be adopted for a linear combination of two geometric distribution. We simulated a sample, of size 1000, from a linear combination of two geometric distributions with true parameter $a = 0.1$, $b = 0.6513$ and $p = 1.1$. We then estimated the parameters from the data set using ten different κ varying from 0.1 to 1. To know the set of estimates which are closest to the true values, one should simply substitute the sets of estimates into (4.27). The "best" set of estimates should have the minimum value of $Var\left[\hat{\boldsymbol{\Theta}}\right]$ when they are put into (4.27). In Figure 5.16, we plot the $Var\left[\hat{\boldsymbol{\Theta}}\right]$ given by the parameter estimates versus κ alongside the plot of $Var\left[\hat{\boldsymbol{\Theta}}\right]$ given by the true parameters. Since $\kappa > 0.5$ produced complex estimates, we exclude them from the plots. For a large separation like this, we know that the best κ is 0.1. Excitingly, when we substitute the estimates given by $\kappa = 0.1$ into (4.27), the resulted $Var\left[\hat{\boldsymbol{\Theta}}\right]$ is the smallest among all the κ considered. From these plots we can confirm that, in practice, the best estimates will have the lowest $Var\left[\hat{\boldsymbol{\Theta}}\right]$ when they are put into (4.27).

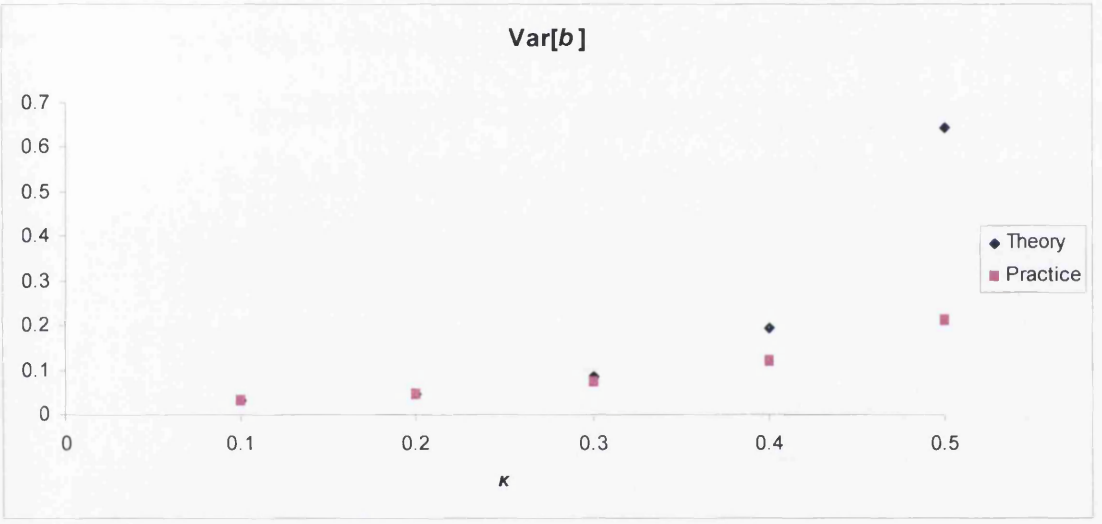
In this subsection, we have seen that the method of rising factorial fractional moments is an ideal method for estimating the parameters from a linear combination of two geometric distributions. We also know that the best κ is generally small (≤ 0.3) and it decreases with the degree of separation between the components. In practice, when r is not known, we should fit a raw sample with different κ and choose the set of estimates that minimises $Var\left[\hat{\boldsymbol{\Theta}}\right]$. *Compared to the MLE via the EM algorithm, this method definitely improves the identification of a linear combination of two geometric distributions.*

The Method of Attenuated Rising Factorial Fractional Moments

Next, we study the performance of the method of attenuated rising factorial fractional moments, studied in Section 4.4, in estimating a linear combination of two geometric distributions from Tables 5.44, 5.45 and 5.46. Like before, we consider 100 combinations of fraction κ and attenuation c, where $\kappa = (0.1, 0.2, ..., 1)$ and $c = (0.01, 0.02, ..., 0.1)$. For each combination, we simulate 10000 artificial samples arising from a linear combination of two geometric distributions and estimate the parameters according to the procedures from (4.41) to (4.53). The minimum measures of error are presented in these three tables with their counterpart combination of κ and c shown in a bracket.

As seen from Table 5.44, when $r = 2$, $p = 1.5$ and $n_o = 1000$, the best combination of





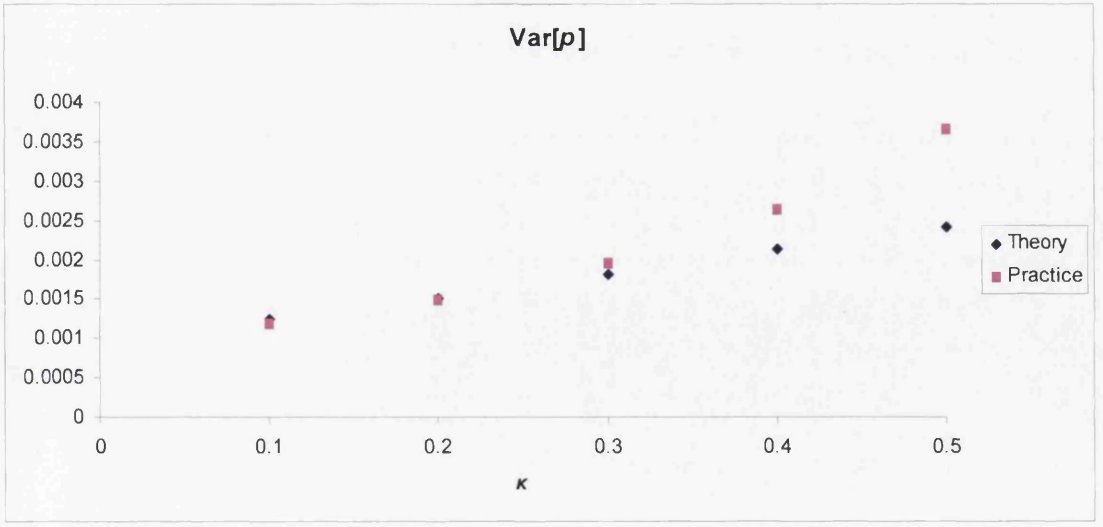


Figure 5.16: Asymptotic variance of the rising factorial fractional moment estimator given by true parameters and parameter estimates versus κ, based on a data set, consisting of 1000 observations, simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$.

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$(\bar{\hat{a}} - a)^2$	0.0005 (0.2, 0.01)	0.0004 (0.1, 0.01)	0.0005 (0.3, 0.01)	0.0003 (0.6, 0.01)	3.66×10^{-13} (0.6, 0.01)
$\left(\bar{\hat{b}} - b\right)^2$	0.0005 (1, 0.09)	9.36×10^{-5} (1, 0.03)	0.0002 (1, 0.07)	0.0001 (1, 0.08)	0.0005 (1, 0.02)
$(\bar{\hat{p}} - p)^2$	0.1116 (0.2, 0.09)	0.0068 (0.3, 0.07)	0.0352 (0.5, 0.1)	0.0061 (0.4, 0.08)	8.83×10^{-7} (0.4, 0.08)
$Var[\hat{a}]$	0.0017 (1, 0.01)	0.0015 (1, 0.01)	0.0014 (1, 0.01)	0.0010 (0.5, 0.01)	5.48×10^{-5} (0.2, 0.01)
$Var\left[\hat{b}\right]$	19 (1, 0.01)	44 (1, 0.02)	51 (0.9, 0.05)	42 (0.9, 0.1)	0.0032 (0.3, 0.01)
$Var[\hat{p}]$	1.1132 (0.5, 0.01)	0.9523 (0.1, 0.01)	0.7043 (0.1, 0.01)	0.8359 (0.1, 0.01)	0.7199 (0.2, 0.07)
$MSE[\hat{a}]$	0.0024 (1, 0.01)	0.0021 (0.1, 0.01)	0.0021 (0.2, 0.01)	0.0013 (0.5, 0.01)	5.54×10^{-5} (0.3, 0.01)
$MSE\left[\hat{b}\right]$	19 (1, 0.01)	44 (1, 0.02)	52 (0.9, 0.05)	43 (0.9, 0.1)	0.004 (0.3, 0.01)
$MSE[\hat{p}]$	1.3621 (0.5, 0.01)	1.1461 (0.1, 0.01)	0.9099 (0.1, 0.01)	1.0078 (0.1, 0.01)	0.7260 (0.2, 0.07)

Table 5.44: Performance of the method of attenuated rising factorial fractional moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 1.5$ for different sample size n_o.

κ and c for both a and b is $\kappa = 0.3$ and $c = 0.01$; whereas the ideal combination for p is $\kappa = 0.2$ and $c = 0.07$, in terms of both the variance and the mean square error. However, the pattern of the best combination for small samples differs with sample size. Generally, when n_o is small, we should use a small fraction ($\kappa \leq 0.5$) and $c = 0.01$ for the estimation of a and p to minimise the mean square errors; whereas for b, the variance is minimised when a large fraction is used, for instance, $\kappa = 0.9$ or 1. In fact, small fractions with a small attenuation do provide reasonable estimates for small samples. However, their $Var\left[\hat{b}\right]$ are large because of a few extremely large estimates of b due to poor estimates of y.

For $r = 5$ and $p = 1.1$, as shown in Table 5.45, the best combination for a, in terms of the mean square error, is a small fraction $\kappa \leq 0.4$ and a small attenuation $0.01 \leq c \leq 0.03$; whereas for b, we should use a large fraction κ being either 0.9 or 1 and a small attenuation $0.01 \leq c \leq 0.04$ for small samples ($n_o \leq 50$) and $\kappa = 0.1$ and $c = 0.01$ for large samples with $n_o = 1000$. For the estimation of p, we should use a small fraction ($\kappa = 0.1$ or 0.2) with an attenuation $c = 0.01$ for small samples and $\kappa = 0.7$ with $c = 0.06$ when the sample size is $n_o = 1000$.

From Table 5.46, we can see that when $r = 10$ and $p = 1.1$, in general, the best combination for a is $\kappa = 0.1$ and $c = 0.01$. We also note that, even when the sample size is as small as $n_o = 10$, the bias of $\hat{a}$ is very small and it decreases with n_o. For b, the best combination is κ being 0.9 or 1 and a small c being 0.01 or 0.02 when n_o is small; whereas

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\bar{\hat{a}} - a\right)^2$	0.0001 (0.2, 0.01)	0.0001 (0.1, 0.01)	0.0001 (0.1, 0.01)	0.0001 (0.2, 0.01)	3.38×10^{-9} (0.2, 0.1)
$\left(\bar{\hat{b}} - b\right)^2$	0.0135 (1, 0.09)	0.0053 (1, 0.09)	0.0011 (1, 0.09)	0.0286 (1, 0.05)	0.0023 (0.1, 0.01)
$\left(\bar{\hat{p}} - p\right)^2$	7.82×10^{-6} (0.1, 0.09)	2.66×10^{-5} (0.2, 0.06)	4.54×10^{-5} (1, 0.1)	0.0003 (0.5, 0.05)	0.0013 (0.1, 0.09)
$Var\left[\hat{a}\right]$	0.0021 (1, 0.01)	0.0018 (1, 0.01)	0.0017 (1, 0.01)	0.0014 (0.2, 0.01)	3.11×10^{-5} (0.4, 0.03)
$Var\left[\hat{b}\right]$	108 (1, 0.04)	101 (0.9, 0.02)	27 (0.9, 0.01)	49 (1, 0.02)	0.0542 (0.1, 0.01)
$Var\left[\hat{p}\right]$	1.0489 (0.2, 0.01)	0.6558 (0.1, 0.01)	0.7187 (0.2, 0.01)	0.5071 (0.1, 0.01)	0.0427 (0.7, 0.06)
$MSE\left[\hat{a}\right]$	0.0025 (1, 0.01)	0.0022 (0.1, 0.01)	0.0021 (0.1, 0.01)	0.0015 (0.2, 0.01)	3.25×10^{-5} (0.4, 0.03)
$MSE\left[\hat{b}\right]$	108 (1, 0.04)	101 (0.9, 0.02)	27 (0.9, 0.01)	49 (1, 0.02)	0.0565 (0.1, 0.01)
$MSE\left[\hat{p}\right]$	1.0602 (0.2, 0.01)	0.6660 (0.1, 0.01)	0.7310 (0.2, 0.01)	0.5159 (0.1, 0.01)	0.0449 (0.7, 0.06)

Table 5.45: Performance of the method of attenuated rising factorial fractional moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 1.1$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$\left(\bar{\hat{a}} - a\right)^2$	2.07×10^{-5} (0.2, 0.01)	1.05×10^{-5} (0.1, 0.03)	6.20×10^{-6} (0.1, 0.01)	7.39×10^{-10} (0.2, 0.05)	1.87×10^{-7} (0.4, 0.08)
$\left(\bar{\hat{b}} - b\right)^2$	0.0239 (0.9, 0.02)	0.0247 (1, 0.02)	0.0557 (0.9, 0.02)	0.0326 (1, 0.03)	8.19×10^{-5} (0.1, 0.1)
$\left(\bar{\hat{p}} - p\right)^2$	2.41×10^{-8} (0.6, 0.09)	1.85×10^{-7} (0.6, 0.03)	6.58×10^{-7} (0.1, 0.02)	7.82×10^{-5} (0.9, 0.01)	3.55×10^{-5} (0.4, 0.08)
$Var\left[\hat{a}\right]$	0.0022 (1, 0.01)	0.0020 (1, 0.01)	0.0017 (0.1, 0.01)	0.0008 (0.1, 0.01)	1.37×10^{-5} (0.1, 0.01)
$Var\left[\hat{b}\right]$	56 (0.9, 0.02)	33 (1, 0.01)	127 (0.9, 0.02)	133 (0.9, 0.02)	0.0234 (0.1, 0.09)
$Var\left[\hat{p}\right]$	0.8113 (0.5, 0.02)	0.6939 (0.2, 0.04)	0.5713 (0.2, 0.02)	0.4707 (0.3, 0.03)	0.0012 (0.1, 0.03)
$MSE\left[\hat{a}\right]$	0.0026 (1, 0.01)	0.0021 (0.1, 0.01)	0.0017 (0.1, 0.01)	0.0008 (0.1, 0.01)	1.39×10^{-5} (0.1, 0.01)
$MSE\left[\hat{b}\right]$	56 (0.9, 0.02)	34 (1, 0.01)	127 (0.9, 0.02)	133 (0.9, 0.02)	0.0235 (0.1, 0.09)
$MSE\left[\hat{p}\right]$	0.8149 (0.5, 0.02)	0.6939 (0.2, 0.04)	0.5714 (0.2, 0.02)	0.4749 (0.3, 0.07)	0.0012 (0.1, 0.03)

Table 5.46: Performance of the method of attenuated rising factorial fractional moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$ for different sample size n_o.

r	p	Theoretical Optimal κ and c	Theoretical $Var[\hat{a}]$	Practical Optimal κ and c	Practical $Var[\hat{a}]$
2	1.5	$(0.5259, 0.0133)$	8.65×10^{-5} (5.72×10^{-5})	$(0.2, 0.01)$	5.48×10^{-5}
5	1.1	$(0.1340, 0.0079)$	2.01×10^{-5} (3.91×10^{-5})	$(0.4, 0.03)$	3.11×10^{-5}
10	1.1	$(0.0977, 0.0124)$	1.33×10^{-5} (1.42×10^{-5})	$(0.1, 0.01)$	1.37×10^{-5}

Table 5.47: Theoretical and simulated minimum variance of the attenuated rising factorial fractional moment estimator $\hat{a}$ given by the optimal κ for a linear combination of two geometric distributions with fixed $a = 0.1$ and $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

for large samples with $n_o = 1000$, the ideal combination for b is $\kappa = 0.1$ and $c = 0.09$. Generally, for the estimation of p, we should use a small fraction $\kappa \leq 0.5$ and a small attenuation $0.02 \leq c \leq 0.07$.

Unlike the positive mixture, we do not find good conformity between the practical and theoretical variance of estimator suggested by (4.54), as observed in Tables 5.47, 5.48 and 5.49. We used the suggested combination of κ and c to estimate another 10000 simulated data sets and the yielded variances of the estimators are presented in brackets underneath the theoretical values in these tables. For a, the $Var[\hat{a}]$ given by the suggested combination is close to, but still slightly lower than the minimum variances we obtained in our previous simulation experiments. For b, the suggested combination, $\kappa = 0.0365$ and $c = 0.0567$, did return estimates of $\hat{b}$ with a marginally smaller $Var\left[\hat{b}\right] = 0.0227$, compared to the variance ($Var\left[\hat{b}\right] = 0.0234$) given by $\kappa = 0.1$ and $c = 0.09$ in our previous estimation. For p, the suggested combination returned higher variances compared to the ones we obtained earlier, except for $r = 10$. From these three tables, we understand that a large range of the combination of κ and c will return estimates with low variances. When the separation between the two components increases, we should use a smaller fraction with an attenuation in order to achieve plausible estimates.

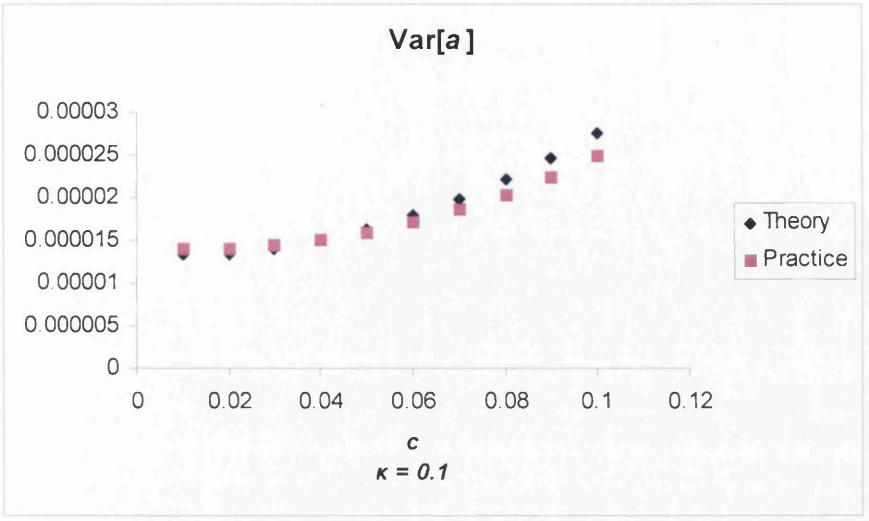
In practice, we do not know the degree of separation between the components, so we need a way to choose a good combination of κ and c. We simulated an artificial sample, consisting of 1000 observations, arising from a linear combination of two geometric distributions with true parameters $a = 0.1$, $b = 0.6513$ and $p = 1.1$. We then estimated the parameters from the sample with ten different combinations of κ and c, where κ is fixed at 0.1 and c varies from 0.01 to 0.1. We substituted the ten sets of estimates into (4.54) and plot the resulted $Var\left[\hat{\Theta}\right]$ versus c alongside the theoretical values of $Var\left[\hat{\Theta}\right]$ given by the true parameters. We can see that the conformity between the theory and practice is satisfactory. Therefore, given a raw sample, one should estimate the parameters with a number of different combinations of κ and c, and substitute the different sets of estimates into (4.54). They should then choose the estimates that minimise $Var\left[\hat{\Theta}\right]$.

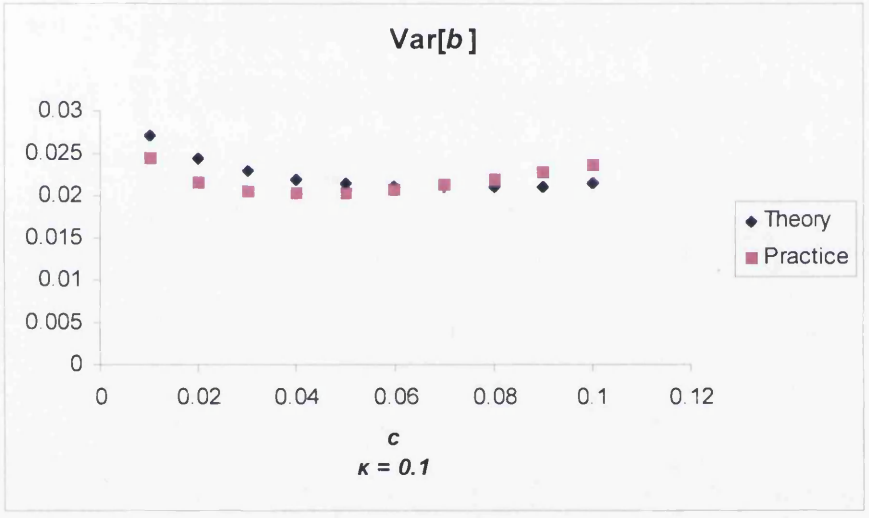
r	p	Theoretical Optimal κ and c	Theoretical $Var\left[\hat{b}\right]$	Practical Optimal κ and c	Practical $Var\left[\hat{b}\right]$
2	1.5	(0.5303, 0.0156)	0.0035 (0.0039)	(0.3, 0.01)	0.0032
5	1.1	(0.1939, 0.0249)	0.0293 (0.0905)	(0.1, 0.01)	0.0542
10	1.1	(0.0365, 0.0567)	0.02049 (0.0227)	(0.1, 0.09)	0.0234

Table 5.48: Theoretical and simulated minimum variance of the attenuated rising factorial fractional moment estimator $\hat{b}$ given by the optimal κ for a linear combination of two geometric distributions with fixed $a = 0.1$ and $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.

r	p	Theoretical Optimal κ and c	Theoretical $Var\left[\hat{p}\right]$	Practical Optimal κ and c	Practical $Var\left[\hat{p}\right]$
2	1.5	(0.5284, 0.0145)	0.2001 (1.2985)	(0.2, 0.07)	0.7199
5	1.1	(0.1604, 0.0173)	0.0037 (0.2386)	(0.7, 0.06)	0.0427
10	1.1	(0.0819, 0.0637)	0.0009 (0.0013)	(0.1, 0.03)	0.0012

Table 5.49: Theoretical and simulated minimum variance of the attenuated rising factorial fractional moment estimator $\hat{p}$ given by the optimal κ for a linear combination of two geometric distributions with fixed $a = 0.1$ and $n_o = 1000$ and various r and p. Simulated figures are based on 10000 replications.





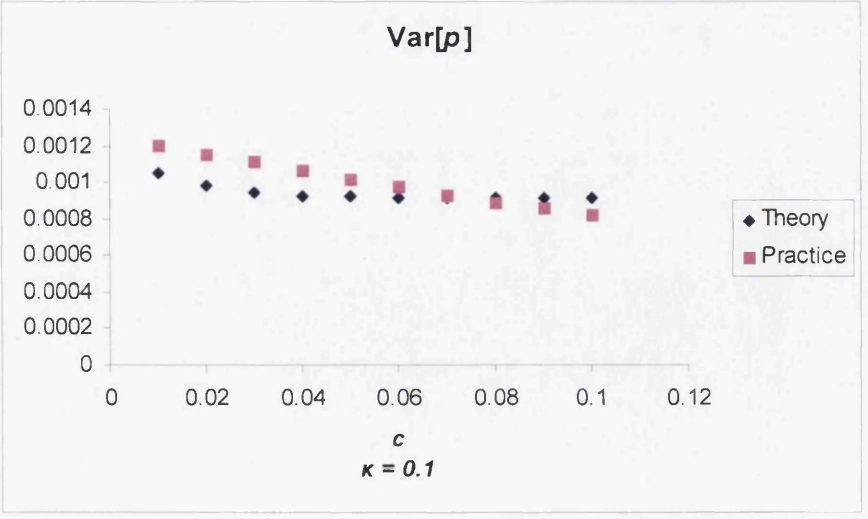


Figure 5.17: Asymptotic variance of the attenuated moment estimator given by true parameters and parameter estimates versus c, based on a data set, consisting of 1000 observations, simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$.

The attenuated rising factorial fractional moment estimator is indeed a highly efficient method for the parameter estimation of a linear combination of two geometric distributions. It is able to identify a linear combination rather than a positive mixture even when the number of observations in a data set is limited ($\hat{p}$ are lowly biased for small n_o as shown in Tables 5.44 to 5.46). Generally, the estimates given by this method have small bias and small variances, especially when the sample size is large. The main drawback of this method is that we may obtain estimates in complex forms, and the estimate of b can be extremely large when a small fraction is used on small samples. Apart from these drawbacks, the method of attenuated rising factorial fractional moments provides estimates of parameters with the best goodness of fit compared to the other methods, as we shall demonstrate in the upcoming section.

The Method of Appell Moments

Finally, we discuss the performance of the method based on double Appell sequences in estimating the parameters from a linear combination of two geometric distributions. Like before, we consider the Kronecker sequences, as explained in Example 4 of Section 4.5.1, and follow the estimation procedures from (4.91) to (4.95).

Tables 5.50, 5.51 and 5.52 present the estimation results of linear combinations of two geometric distributions with different separation $r = 2$, 5 and 10 respectively. When the sample size is small, this method recognises the distribution as a positive mixture, with a large proportion on the first component. Indeed, the estimation is not satisfactory because most of the estimates of b have negative values. We note, in the tables, that the average of $\hat{b}$ for small samples is negative in most cases. In general, the variance of $\hat{b}$ is large for all r, even when samples are large, $n_o = 1000$.

When the sample size is small, $E[\hat{p}]$ does not exceed 1. In other words, the fitted distribution is a positive mixture when data is scarce. As a result, the bias of $\hat{p}$ is considerably large for small samples. We also note the large variance of $\hat{p}$ in Tables 5.50 and 5.52. When the sample size is increased to 1000, we see a great improvement in the estimation of p with a much lower bias[2].

We investigated the estimation results and found that $Var\left[\hat{\delta}_2\right]$ and $Var\left[\hat{\delta}_3\right]$ are extremely large even when the sample size is as large as 1000. For instance, when $r = 10$, $p = 1.1$ and $n_o = 1000$, $Var\left[\hat{\delta}_2\right]$ is 30469 and $Var\left[\hat{\delta}_3\right]$ is 18253922. Due to the highly inconsistent raw moments, the estimates of b and p are poor even for large sample sizes.

Therefore, we conclude that the method using the Appell sequences is not suitable as it is very likely to provide unreasonable estimates of b and p. This method, like the MLE, is not to be favoured for the estimation problem of a linear combination of two geometric distributions.

$r = 2$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1186	0.1094	0.1045	0.0887	0.0988
$E\left[\hat{b}\right]$	−0.2512	−0.0274	−0.3178	−0.2168	0.2637
$E[\hat{p}]$	0.9107	0.9285	0.9290	0.8835	1.3892
$(\bar{\hat{a}} - a)^2$	0.0004	9.00×10^{-5}	2.00×10^{-5}	0.0001	1.44×10^{-6}
$\left(\bar{\hat{b}} - b\right)^2$	0.1947	0.0473	0.2579	0.1655	0.0054
$(\bar{\hat{p}} - p)^2$	0.3473	0.3266	0.3261	0.3801	0.0123
$Var[\hat{a}]$	0.0018	0.0011	0.0007	0.0005	6.54×10^{-5}
$Var\left[\hat{b}\right]$	563	520	177	157	52
$Var[\hat{p}]$	1.5757	0.0917	1.9881	0.1147	1.6279
$MSE[\hat{a}]$	0.0021	0.0012	0.0008	0.0007	6.69×10^{-5}
$MSE\left[\hat{b}\right]$	563	520	177	157	52
$MSE[\hat{p}]$	1.9230	0.4183	2	0.4948	1.6402

Table 5.50: Performance of the method of Appell moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 1.5$ for different sample size n_o.

$r = 5$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1297	0.1175	0.1118	0.0948	0.1002
$E\left[\hat{b}\right]$	−0.9500	0.0858	−0.0352	−0.0826	0.1633
$E[\hat{p}]$	0.8959	0.9079	0.9166	0.8602	1.0935
$(\bar{\hat{a}} - a)^2$	0.0009	0.0003	0.0001	3.00×10^{-5}	2.34×10^{-8}
$\left(\bar{\hat{b}} - b\right)^2$	1.8484	0.1048	0.1978	0.2421	0.0606
$(\bar{\hat{p}} - p)^2$	0.0417	0.0369	0.0337	0.0575	4.22×10^{-5}
$Var[\hat{a}]$	0.0024	0.0015	0.0010	0.0007	0.0001
$Var\left[\hat{b}\right]$	822	340	266	172	23
$Var[\hat{p}]$	1.0063	0.2283	0.0663	0.0935	0.2660
$MSE[\hat{a}]$	0.0033	0.0018	0.0012	0.0007	0.0001
$MSE\left[\hat{b}\right]$	824	340	266	172	23
$MSE[\hat{p}]$	1.0480	0.2652	0.0999	0.1510	0.2660

Table 5.51: Performance of the method of Appell moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 1.1$ for different sample size n_o.

$r = 10$	Simulated Value				
n_o	10	15	20	50	1000
$E[\hat{a}]$	0.1308	0.1181	0.1120	0.0955	0.1005
$E[\hat{b}]$	−0.2515	−0.3242	−0.0570	−0.0205	0.2822
$E[\hat{p}]$	0.8487	0.8752	0.8908	0.8357	1.0880
$(\hat{a}-a)^2$	0.0010	0.0003	0.0001	2.07×10^{-5}	2.85×10^{-7}
$(\hat{b}-b)^2$	0.8151	0.9516	0.5017	0.4514	0.1363
$(\hat{p}-p)^2$	0.0632	0.0505	0.0438	0.0699	0.0001
$Var[\hat{a}]$	0.0025	0.0015	0.0011	0.0007	0.0001
$Var[\hat{b}]$	299	278	219	175	49
$Var[\hat{p}]$	6	1.3619	1.8150	1.8577	0.3887
$MSE[\hat{a}]$	0.0034	0.0018	0.0012	0.00069	0.0001
$MSE[\hat{b}]$	300	279	220	175	49
$MSE[\hat{p}]$	6	1.4124	1.8588	1.9276	0.3888

Table 5.52: Performance of the method of Appell moments for 10000 data sets simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$ for different sample size n_o.

5.2.4 Comparison of Estimation Methods

We have presented the estimation results of four methods: the MLE via the EM algorithm (ML), the rising factorial fractional moments (FM), the attenuated rising factorial fractional moments (AM) and the method based on Appell sequences (AP). It is clear that the MLE and the AP are not plausible for estimating the parameters in a linear combination of two geometric distributions. The MLE fits a positive mixture distribution for small samples and a single distribution for large samples; when $n_o = 1000$, $\hat{a}$ and $\hat{b}$ are close to each other, as seen in Figure 5.18. In this figure, we compare the scatter plots of $\hat{b}$ versus $\hat{a}$ given by the MLE and the AM for three degrees of separation $r = 2$, 5 and 10. The MLE of $\hat{a}$ and $\hat{b}$ are positively correlated and have similar values in each plot; we can see that none of the 10000 estimates of b are close to the true value but they are all very similar to $\hat{a}$. On the other hand, the AM returns more reasonable estimates of b. When $r = 2$, we observe a few outliers of $\hat{b}$ with extremely large values in the scatter plot; these outliers normally make the variance of estimator $\hat{b}$ large. From these plots, we discover that the AM outperforms the MLE.

On the other hand, the AP is highly likely to provide negative estimates of b for samples of small sizes; for samples of large sizes, the estimates of b are both highly biased and with large variance. Since both the AP and the MLE are not favoured, we only focus on the FM and the AM to conclude on the best method.

Tables 5.53 shows the performance of the four methods in estimating 10000 data sets, each consisting 1000 observations, arising from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 1.5$. In terms of the bias, the AM is better

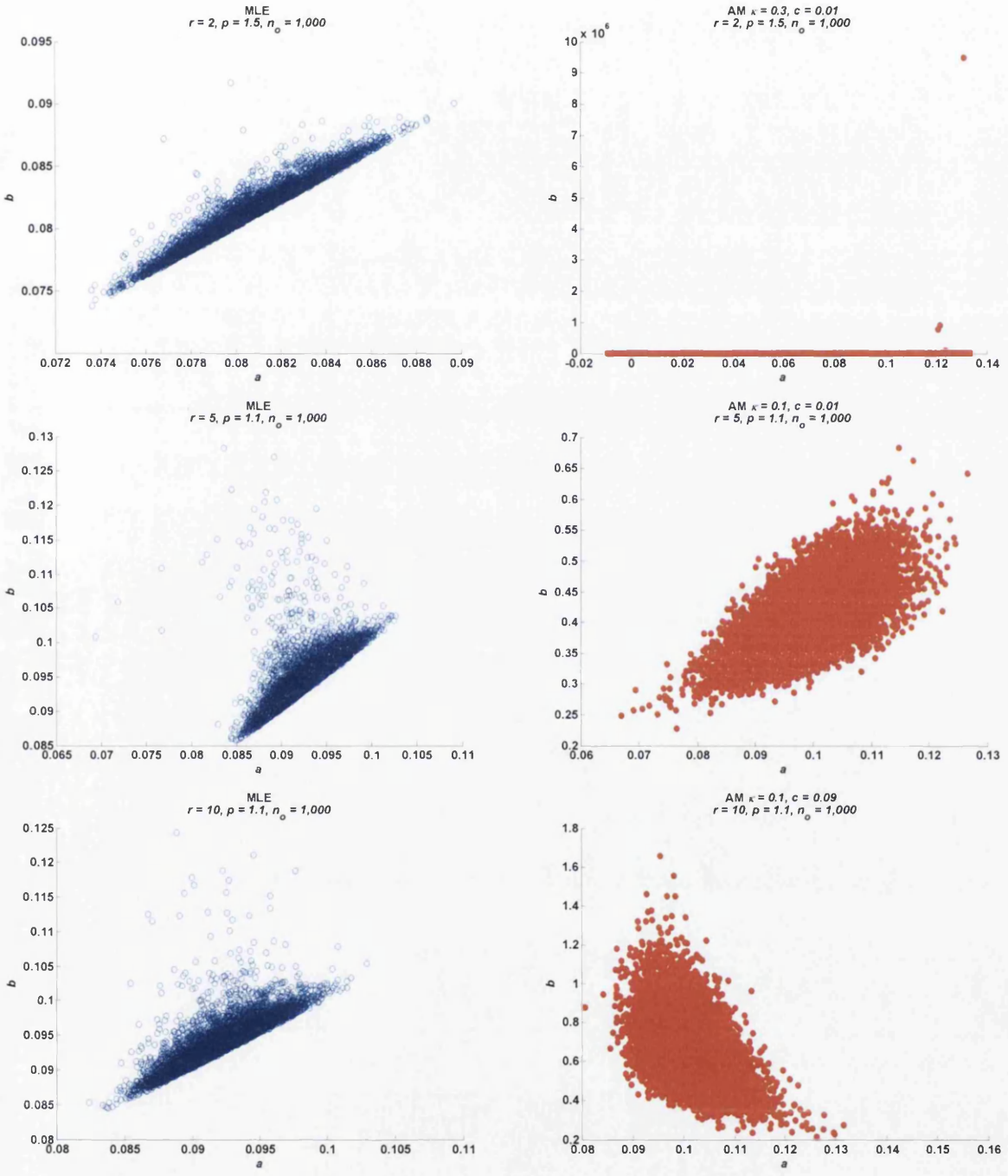


Figure 5.18: Scatter plots of b versus a: Comparison of the MLE and attenuated moment estimator for a linear combination of two geometric distributions with different r and p.

$r = 2$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
ML	0.0004	0.0118	0.6430	4.56×10^{-6}	4.79×10^{-6}	3.25×10^{-5}
FM	7.55×10^{-10}	0.0003	0.01426	5.46×10^{-5}	0.0034	1.1399
AM	3.66×10^{-13}	0.0005	8.83×10^{-7}	5.48×10^{-5}	0.0032	0.7199
AP	1.44×10^{-6}	0.0054	0.0123	6.54×10^{-5}	52	1.6279

Table 5.53: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.19$ and $p = 1.5$.

$r = 5$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
ML	5.30×10^{-5}	0.0990	0.0654	7.16×10^{-6}	1.09×10^{-5}	0.0003
FM	1.43×10^{-6}	0.0105	0.0023	2.98×10^{-5}	0.0786	0.0499
AM	3.38×10^{-9}	0.0023	0.0013	3.11×10^{-5}	0.0542	0.0427
AP	2.34×10^{-8}	0.0606	4.22×10^{-5}	0.0001	23	0.2660

Table 5.54: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.4095$ and $p = 1.1$.

in estimating a and p; The FM has marginally smaller variance of $\hat{a}$ and $\hat{b}$ in this case. However, its variance of $\hat{p}$ is significantly larger than the one given by the AM.

We now move on to the study of these estimators performance on distributions with medium separation between the components, $r = 5$ and $p = 1.1$ in Table 5.54. The AM is better than the FM for estimating all parameters because it has lower bias^2. In terms of the variance, the AM is more efficient for the estimation b and p than the FM. The variance of $\hat{a}$ given by the FM is marginally smaller than the one given by the AM.

Finally, we study the performance of these methods on distributions with well-separated components, $r = 10$ and $p = 1.1$ in Table 5.55. Once again, the AM estimators have lower bias and lower variances. Excitingly, the performance of the FM in this case is comparable to the AM because its variances of estimators are only slightly larger than the ones given by the AM.

In conclusion, for a linear combination of two geometric distributions, the method of

$r = 10$	Bias^2			Variance		
$n_o = 1000$	a	b	p	a	b	p
ML	6.14×10^{-5}	0.3105	0.0265	6.44×10^{-6}	9.00×10^{-6}	3.73×10^{-5}
FM	2.81×10^{-7}	0.0012	5.91×10^{-5}	1.41×10^{-5}	0.0215	0.0012
AM	1.87×10^{-7}	8.19×10^{-5}	3.55×10^{-5}	1.37×10^{-5}	0.0227	0.0012
AP	2.85×10^{-7}	0.13625	0.0001	0.0001	49	0.3888

Table 5.55: Performance of different estimation methods for 10000 data sets each consisting of 1000 observations simulated from a linear combination of two geometric distributions with $a = 0.1$, $b = 0.6513$ and $p = 1.1$.

attenuated rising factorial fractional moments stands out from the others as the best method for the estimation problem. For any sample size, the attenuated rising factorial fractional moment estimator has both lowest bias and lowest variance of all three parameters. This method is especially good at estimating p as it has a very low bias^2 and variance compared to the other methods. This means that the method of attenuated rising factorial fractional moments is good at "tracking" negative weights and distinguishing a linear combination from a positive mixture.

The rising factorial fractional moments is comparable to the method of attenuated rising factorial fractional moments when the sample size is large enough. We have seen that its variances of estimators are close to the ones given by the attenuated moment estimator especially when the two components are well separated.

5.2.5 Discussion

As a summary, the MLE via the EM algorithm is not a good method for estimating the parameters in a linear combination of two geometric distributions, even when the sample size is large enough. The famous MLE either fits a positive mixture distribution to a small data set, or it fits a single distribution, with $\hat{a}$ and $\hat{b}$ similar to each other, to a large data set. The best method considered for this estimation problem is obviously the method of attenuated rising factorial fractional moments. It provides estimates with both low bias and small variances. The rising factorial fractional moment estimator is also a plausible method, provided that the sample size is large enough. We have also observed how a negative fraction reduces the variance of the estimator for a distribution with $r = 10$ and $p = 1.1$. The method using Appell moments is clearly a poor method for the parameter estimation of such a distribution. Like the MLE via the EM algorithm, it is not good in recognising that p is larger than 1 for small samples. For large samples, its estimate of b is highly biased compared to the other methods.

5.3 Summary

In this chapter we addressed the problem of sampling from a probability density/mass function which is a linear combination of components. We focussed on a linear combination of two exponential distributions and its discrete analogue, geometric distributions. The estimation of the parameters in these distributions is straightforward; we simply employ the estimation methods studied in Chapters 3 and 4 because the PDF/PMF is identical to the positive mixture.

For a linear combination of two exponential distributions, the MLE via the EM algorithm fails to fit a reasonable distribution to the simulated samples. It either fits a positive mixture to a small sample, or a single distribution (a and b have similar values) to a large sample. We have also shown that the MLE is not sensitive to the starting values. Even when we started the EM iterative process with true values, the updated estimate $\hat{b}^{(k)}$ falls to a value near

to $\hat{a}^{(k)}$ when the number of iteration increases. The performance of the fractional moment estimator is satisfactory, especially for a sample of large size. The attenuated moment estimator is found to be the best method. It almost always identifies that a distribution is a linear combination rather than a positive mixture when the number of observations in a data set is limited. Not to mention that, for large samples, it has the lowest bias and variance for all three parameters in most cases. The Appell moment estimator is not favoured because it provides estimates with relatively larger variances compared to methods like the fractional moment estimator and the attenuated moment estimator. The method of order statistics is actually good in estimating p. However, it can sometimes return extremely large estimates of b and hence it is outperformed by the other methods.

For the discrete analogue, a linear combination of two geometric distributions, we have seen that the only two plausible methods are the method of rising factorial fractional moments and the method of attenuated rising factorial fractional moments. The attenuated moment estimator has the lowest bias and the smallest variances in most cases, even for small samples. The fractional moment estimator has small variances which are close to the ones given by the attenuated moment estimator when the sample size is large enough.

We have seen in Chapters 3 and 4 that although the performance of the generalised method of moments we investigated are comparable to the MLE, but the MLE is still asymptotically the best method for estimating the parameters from a mixture distribution. In this chapter, we have shown that the MLE via the EM algorithm is outperformed by the fractional moment estimator and the attenuated moment estimator for the parameter estimation of a linear combination of distributions. Of course, the potential problems with moment estimator, namely the possibility of getting complex or negative estimates from moment estimator, still apply.

Chapter 6

Modelling the Incubation Period of Prion Diseases

We reviewed the history of statistical models for the incubation period of an infectious disease in Chapter 2. The standard model for the time between exposure and the manifestation of disease for many cases has been shown to be well approximated by the lognormal distribution. We also introduced the central role that incubation period plays in the experimental investigation of prion diseases, such as scrapie, BSE and CJD. These diseases are thought to be caused by an infectious protein (PrP^{Sc}) rather than a conventional virus, bacteria or parasite. Since the infectious agent is not fully characterised, the incubation period is used as an indicator of virulence (short incubation periods) and to characterise strains of prion disease (by repeatable patterns of incubation period in specific rodent model experimental systems). In this chapter we seek an appropriate model for the prion incubation period.

We test the model on a serial passage experiment, another cornerstone of prion research (outlined in Chapter 1), which highlights key features of the prion incubation period that must be captured by a satisfactory model. We consider a range of standard statistical models used for waiting times. However, since the model needs to be flexible enough to fit characteristics that can change with dose, and generation of passage or strain, we also fit, for the first time, a set of mixture models to the prion disease incubation period.

6.1 Experimental Data

In this chapter, we provide a suitable statistical model for the incubation period of mice which are exposed to the infectious prion protein. In the experiment, a mouse is exposed, orally, to infectious material derived from a case of chronic wasting disease in deer (a member of the prion disease family with similar characteristics to scrapie). This is the "first passage" of the experiment. After a certain time period, when disease symptoms are manifest, the brain of the infected mouse is taken and an oral dose is prepared for exposure to a number of further second passage mice. The experiments were carried on for five passages. All

T_{g_1}	T_{g_2}	T_{g_3}	T_{g_4}	T_{g_5}
408	129 171	157	87*	131
415	129 233	157	133	138
482	133 338	157	133	138
481	133 338	161	138	157
488	152 424	161	138	157
508	237 438	161	167	
521	228	167		
593	217	174		
596	184			
504	169			

mice showed disease, apart from one that died early from other causes, and the incubation periods were recorded, as shown in Table 6.1, where T_{g_i} denotes incubation period from the i^{th} passage, $i = 1, ..., 5$. It is obvious that the animals from third passage onwards tend to have shorter incubation period than the animals from the first passage. It took an average of 500 days for the first passage mice to show disease. The range of incubation period of second passage mice is much larger (range $= (129, 438)$, mean $= 228$) than the others. The average incubation period for third, fourth and fifth passage mice are 162, 141 and 144 respectively. The histogram of the overall incubation period are shown in Figure 6.1; Figure 6.2 presents the histograms of incubation period from each passage, it is clear that earlier passage mice have longer incubation period.

We investigate the incubation period using a box plot with groupings given by the number of passages, as shown in Figure 6.3. It is now easy to pick up the essential message: incubation periods of first passage mice have the highest median and a low variability compared to the second passage mice. Indeed, the median of incubation period decreases when the number of passages increases. Incubation periods of second passage mice have the highest variability among all passages. Remarkably, after second passage, incubation periods are a lot lower than the first passage.

Our aim is to find a suitable model for the prion disease. We first use statistical tests to check if the null hypothesis that two incubation period data sets from different passage are identical can be rejected; Kolmogorov-Smirnov test and Mann-Whitney test are helpful for such purpose. Having done this, we set a prior distribution for the incubation period of each passage and estimate the parameters using the maximum likelihood estimator. The goodness of fit of the fitted distributions are then examined. Lastly, the biological implications of the fitted models are discussed.

6.2 Non-Parametric Test

In this section, we investigate whether the incubation periods from different passages are drawn from the same distribution. For this purpose, we employ two nonparametric tests:

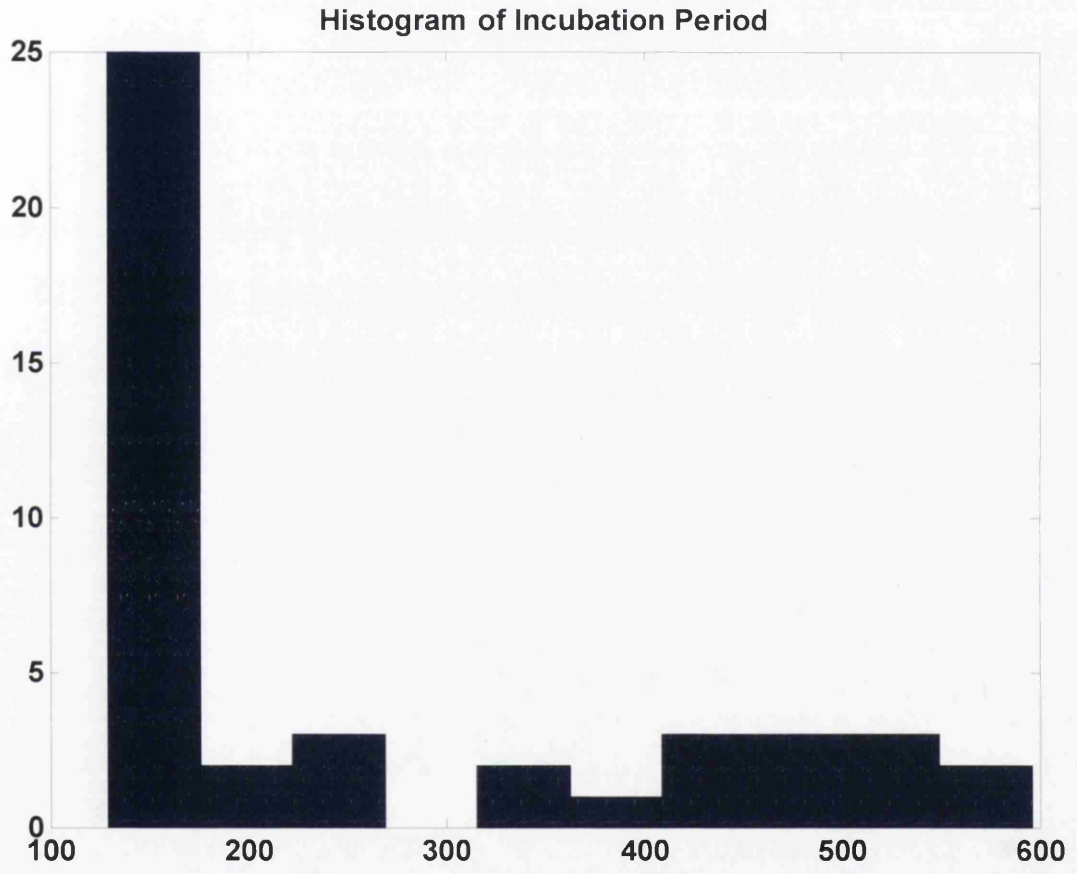


Figure 6.1: Histogram of incubation period (in days) of CWD infected mice.

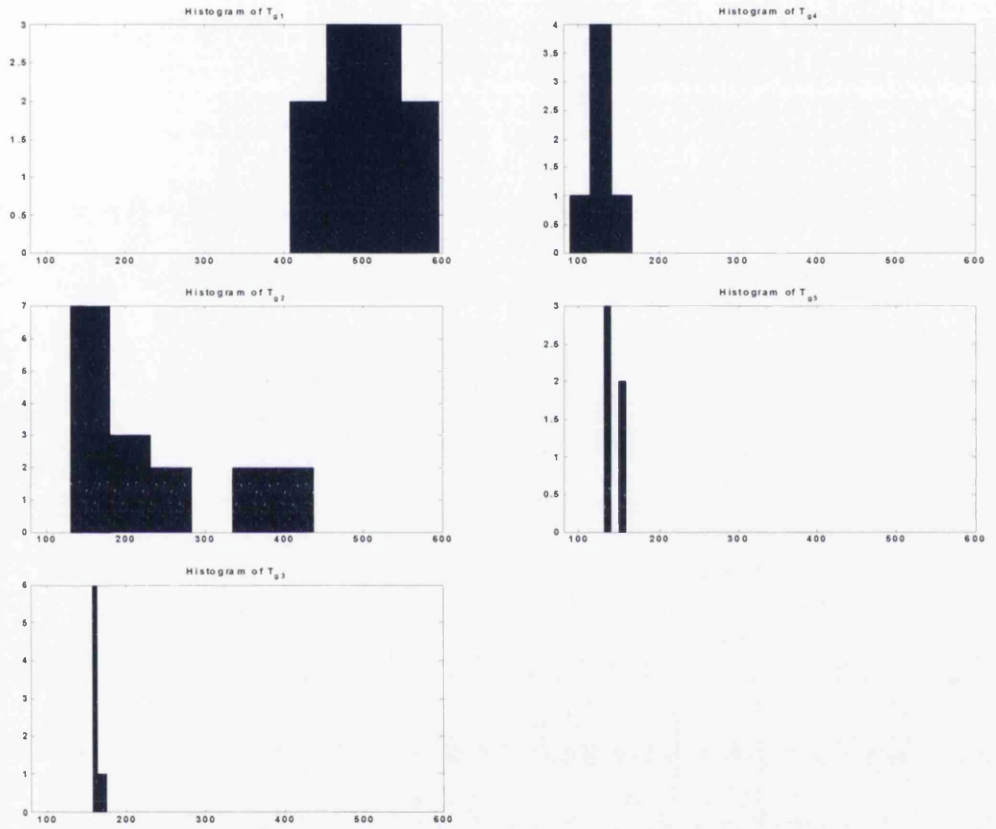


Figure 6.2: Histograms of incubation period (in days) of CWD infected mice from 1^{st} passage (T_{g_1}) to 5^{th} passage (T_{g_5}).

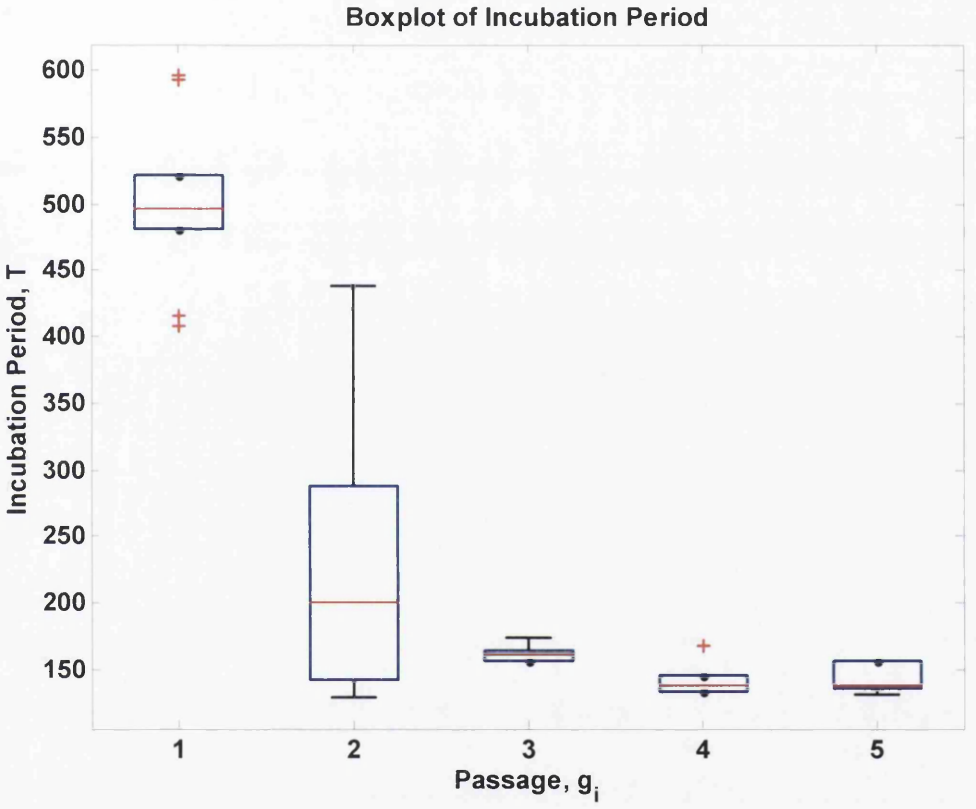


Figure 6.3: Box plot of incubation period (in days) grouped by passage.

the Kolmogorov-Smirnov Test and the Mann-Whitney Test.

6.2.1 Kolmogorov-Smirnov Test

The two-sample KS test is one of the most general non-parametric tests for comparing two samples. It is good at detecting differences in both the locations and the shapes of the distributions. Suppose we have two samples X and Y of length n_1 and n_2. The null hypothesis for this test is that X and Y are drawn from the same continuous distribution. The Kolmogorov-Smirnov test is based on the maximum absolute difference between the observed CDFs for both samples. When this difference is significantly large, the two distributions are considered different. The KS test statistics, KS is given by

$$KS = \max\left[\left|F_1(t) - F_2(t)\right|\right],$$

where $F_1(t)$ is the proportion of X values less than or equal to t and $F_2(t)$ is the proportion of y less than or equal to t.

We apply the KS test on the incubation period data where X is the data vector of the incubation period from passage i and Y is data vector of the incubation period from j^{th} passage. Using SPSS, the results of the KS test for all combinations are shown in Table 6.2. As expected, passage 1 is significantly different from all other passages; the difference between passage 2 and all other passages, except passage 1, is not significant at 5% level. passage 3 is significantly different from passage 4; while the null hypothesis which states that passage 3 and passage 5 are drawn from the same distribution cannot be rejected at 5% level. The difference between passage 4 and passage 5 is not significant.

6.2.2 Mann-Whitney Test

The Mann-Whitney U test is the most popular of the two-independent-sample tests. Suppose we want to test if the the distribution of incubation period from the primary passage animals is significantly different from the one from the second passage animals. We let $X_1, X_2, ..., X_{n_1}$ denote the random sample of size n_1 from the i^{th} passage and let $Y_1, Y_2, ..., Y_{n_2}$ be the random sample of size n_2 from the j^{th} passage. If $F_1(t)$ and $F_2(t)$ are the distribution functions corresponding to i^{th} passage and j^{th} passage, respectively, then the null hypothesis is stated as

$$\begin{aligned} H_0 &: F_1(t) = F_2(t), \\ H_1 &: F_1(t) \neq F_2(t). \end{aligned}$$

Both samples from i^{th} and j^{th} passage are combined into a single ordered sample and ranks are assigned to the sample values from the smallest value to the largest. Let $R(X_i)$ and $R(Y_j)$ denote the rank assigned to X_i and Y_j for all i and j. The Mann-Whitney U statistic

Combined Data	KS	sig_{KS}	U	sig_U
Passage 1 against Passage 2	2.171	0.000	4	0.000
Passage 1 against Passage 3	2.108	0.000	0	0.000
Passage 1 against Passage 4	1.826	0.003	0	0.002
Passage 1 against Passage 5	1.826	0.003	0	0.002
Passage 2 against Passage 3	1.299	0.068	42	0.177
Passage 2 against Passage 4	1.342	0.055	19	0.082
Passage 2 against Passage 5	1.342	0.055	20	0.098
Passage 3 against Passage 4	1.403	0.039	7	0.045
Passage 3 against Passage 5	1.096	0.181	3	0.010
Passage 4 against Passage 5	0.316	1.000	11	0.745

Table 6.2: Kolmogorov-Smirnov test and Mann-Whitney test for incubation period data

is the smaller of U_1 and U_2, where

$$\begin{aligned} U_1 &= \left(\sum_{i=1}^{n_1} R(X_i)\right) - \frac{n_1(n_1+1)}{2}, \\ U_2 &= \left(\sum_{j=1}^{n_2} R(Y_j)\right) - \frac{n_2(n_2+1)}{2}, \\ U &= \min[U_1, U_2]. \end{aligned}$$

Using SPSS, the results of the Mann-Whitney test are shown in Table 6.2. We can observe that the Mann-Whitney test, in general, agrees with the KS test, except for passage 3 versus passage 5. According to the Mann-Whitney test, passage 3 and passage 5 are significantly different from each other at 1% level. We shall now move on to model the incubation period using a parametric approach.

6.3 Fitting the Incubation Period Data of Each Passage with a Single Distribution

From the previous section, we learned that there are significant differences between the length of incubation period data of different passages. For instance, the incubation period data from the first passage is significantly longer than the other passages' incubation period. At this stage, the question we ask is: How is the incubation period of each passage distributed?

Note that the first observation ($t_1 = 87$) of the 4^{th} passage (in Table 6.1) is the only censored data in the sample, for simplicity, we exclude it from the sample. We have fitted the incubation period of each passage with a variety of distributions and found that the normal distribution, gamma distribution and Weibull distribution appeared to provide reasonable fit to the data. The PDFs of these distributions and the MLE of the parameters are summarised in Table 6.7.

Combined Data	$\hat{\mu}$	$\hat{\sigma}$	Mean	Median	Mode	Variance
Passage 1	6.2068	0.1251	500.0085	496.1112	488.4074	3943.4115
Passage 2	5.3448	0.4182	228.6622	209.5160	175.8988	9992.7989
Passage 3	5.0862	0.0360	161.8788	161.7740	161.5644	33.9834
Passage 4	4.9506	0.0954	141.9040	141.2597	139.9799	184.1042
Passage 5	4.9684	0.0829	144.2916	143.7966	142.8118	143.5768

Table 6.3: Fitting every passage's incubation period with a lognormal distribution.

6.3.1 Lognormal Model

The most common distribution for incubation period data is the lognormal, with associated PDF

$$f(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log[t]-\mu)^2}{2\sigma^2}\right],$$

where $t > 0$, $-\infty < \mu < \infty$ and $\sigma > 0$. The parameters μ and σ are estimated, using the MLE, as

$$\hat{\mu} = \frac{1}{n_o}\sum_{i=1}^{n_o} \ln[t_i]$$

and

$$\hat{\sigma} = \sqrt{\frac{1}{n_o}\sum_{i=1}^{n_o} (\ln[t_i]-\hat{\mu})^2}.$$

Having estimated μ and σ for each passage with the MLE, we test the goodness of fit using the KS test. We know that the critical region of the KS test is no longer valid if the parameters are estimated from the data. However, the main purpose for us to undertake this test is to compare the KS distances given by different models. Therefore, the significance values are also included so that we have a rough picture of how the models perform compared to each other. We summarise the estimation results and KS statistics in Table 6.7; from the table, we can see that the goodness of fit for each passage is satisfactory. The estimated mean, median, mode and variance are shown in Table 6.3. Using the estimated parameters, we present the fitted theoretical PDF plots of the lognormal distributions on Figure 6.4. According to the lognormal models, the first passage mice are unlikely to have incubation periods which are less than 300 days. The first passage incubation period is most probable at $T_{g_1} = 488$ days. The variance of the incubation period from the second passage is the largest among all passages. We observe that the range of the second passage incubation period is large and the mode is 175 days. The variance of the third passage incubation period is the smallest among all passages. From the third passage onwards, the characteristics of the incubation periods are similar to each other with a shorter average incubation period and a smaller variance. As seen in Figure 6.4, the distribution of incubation period from every passage is right-skewed (i.e. the right tail of the PDF plot is longer than the left tail and mean > median > mode).

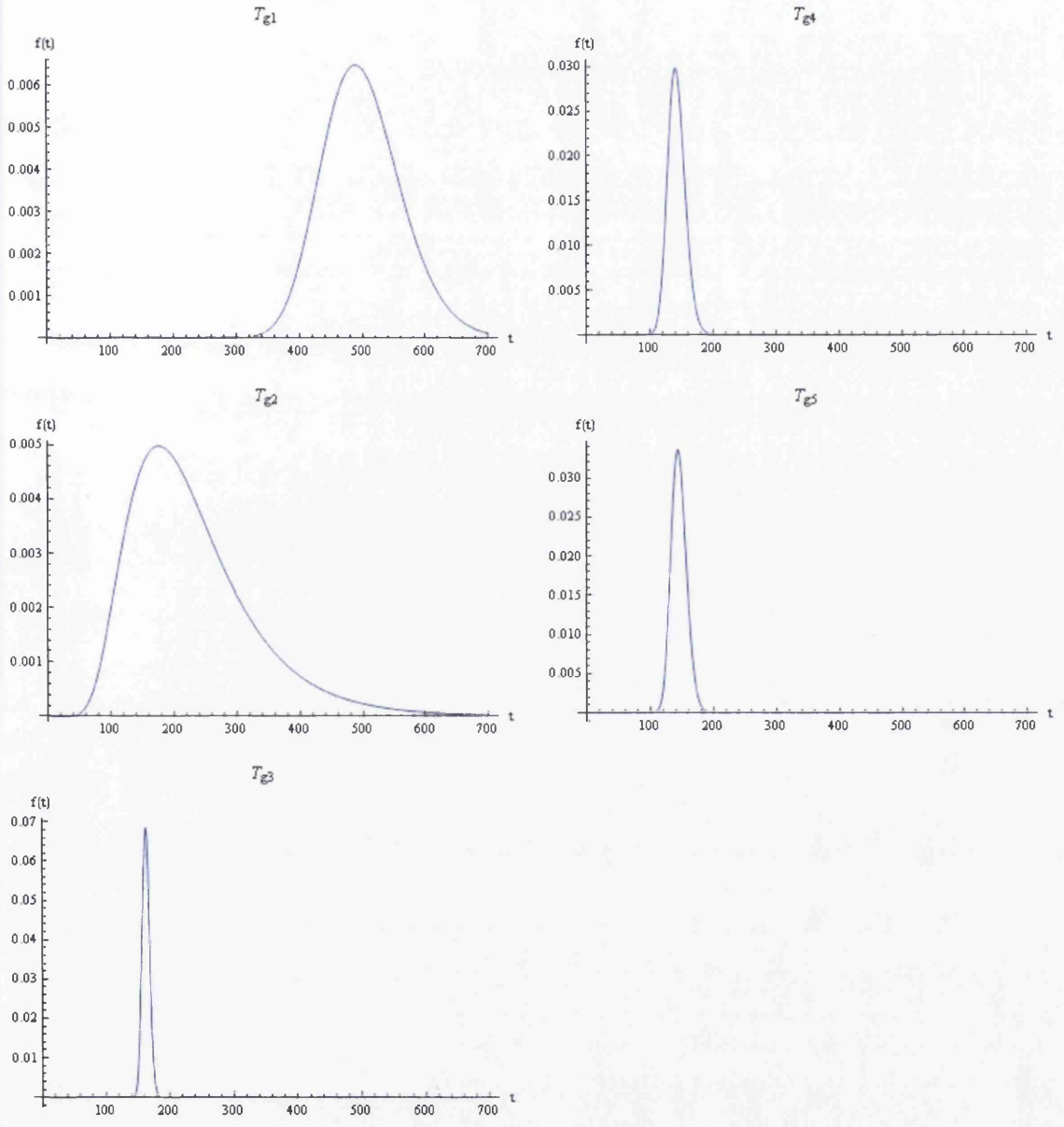


Figure 6.4: The lognormal models for generation 1 to generation 5.

Combined Data	$\hat{\mu}$	$\hat{\sigma}$	Mean = Median = Mode	Variance
Passage 1	499.6000	62.2739	499.6000	3878.0386
Passage 2	228.3125	102.9678	228.3125	10602.3678
Passage 3	161.8750	5.9387	161.8750	35.2682
Passage 4	141.8000	14.3073	141.8000	204.6988
Passage 5	144.2000	12.0291	144.2000	144.6992

Table 6.4: Fitting every passage's incubation period with a normal distribution.

6.3.2 Normal Model

The normal distribution is the most widely used family of distribution in statistics. The PDF of a normal distribution is

$$f(t;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma_1}} \exp\left(-\frac{1}{2}\left[\frac{t_i-\mu_1}{\sigma_1}\right]^2\right),$$

where $t \in \mathbb{R}$ and $\sigma > 0$. The MLE of the parameters are

$$\hat{\mu} = \frac{1}{n_o}\sum_{i=1}^{n_o} t_i$$

and

$$\hat{\sigma} = \sqrt{\frac{1}{n_o}\sum_{i=1}^{n_o}(t_i-\hat{\mu})^2},$$

i.e. $\hat{\mu}$ is given by the sample mean; whereas $\hat{\sigma}$ is estimated by the standard deviation of the data. We fit the incubation data from each passage with a normal distribution and use KS test to check the goodness of fit. The estimation results and some of the descriptive statistics are shown in Table 6.4; while KS statistics are shown in Table 6.7. Figure 6.5 shows the theoretical PDF plots of normal distributions using the estimation results of each passage. The characteristics of the fitted normal distribution for each passage are similar to the ones suggested by the lognormal models in the previous section (compare Figure 6.4 with Figure 6.5). We know that the mean, median and mode of a normal distribution have the same value. Therefore, the normal models suggest that each passage mice are most probable to have incubation period as the average period of the data, i.e. when $T_{g_1} = 499$ days, $T_{g_2} = 228$ days, $T_{g_3} = 161$ days, $T_{g_4} = 141$ days and $T_{g_5} = 144$ days for each passage respectively (see Table 6.4). However, the fitted normal model for the incubation periods of second passage mice allows a negative incubation period, which is of course unrealistic in practice.

6.3.3 Gamma Model

It appears that gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\theta > 0$ also fits well to the incubation period data. A variable T is gamma distributed, denoted

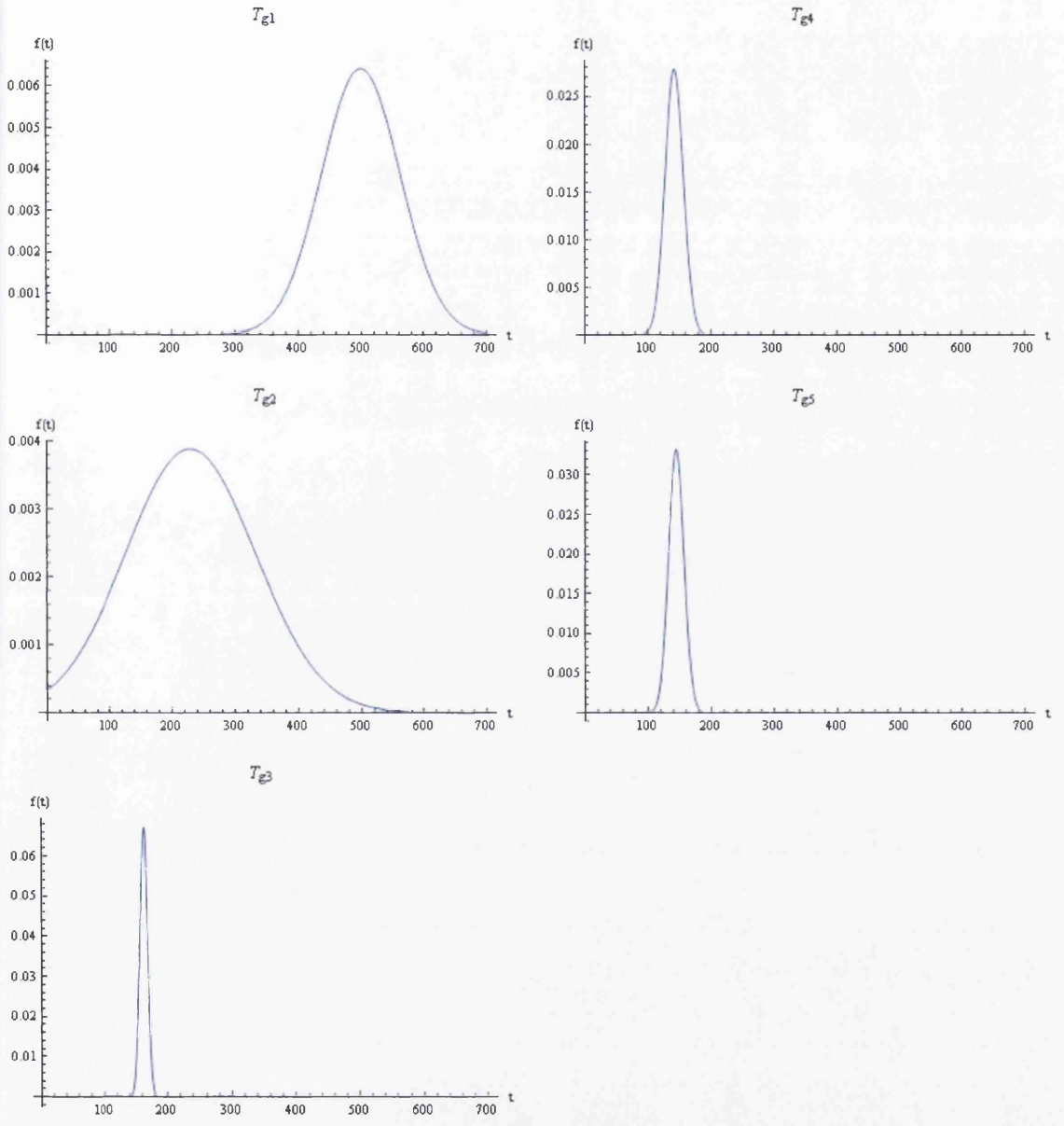


Figure 6.5: The normal models for generation 1 to generation 5.

$T \sim \Gamma(\alpha, \theta)$, if its density function is

$$f(t; \alpha, \theta) = \frac{\theta^{\alpha}}{\Gamma(\alpha)} \exp(-\theta t)\, t^{\alpha-1}$$

where $t > 0$. The shape of a gamma distribution (either bell-shaped or L-shaped) depends on the value of α; this flexibility has enabled the gamma distribution to capture many specific properties of real data.

Suppose we have a sample of n_o iid observations which are gamma distributed. The log-likelihood function, denoted by $l(\boldsymbol{\Theta})$ is

$$l(\boldsymbol{\Theta}) = (\alpha - 1) \sum_{i=1}^{n_o} \log t_i - \theta \sum_{i=1}^{n_o} t_i + \alpha n_o \log \theta - n_o \log \Gamma(\alpha). \tag{6.1}$$

Setting the score function with respect to θ to zero yields the ML estimate of θ:

$$\hat{\theta} = \frac{\alpha n_o}{\sum_{i=1}^{n_o} t_i}. \tag{6.2}$$

Substituting (6.2) into (6.1), the log-likelihood function is in the form of

$$l(\boldsymbol{\Theta}) = (\alpha - 1) \sum_{i=1}^{n_o} \log t_i - \alpha n_o + \alpha n_o \log\left(\frac{\alpha n_o}{\sum_{i=1}^{n_o} t_i}\right) - n_o \log \Gamma(\alpha). \tag{6.3}$$

By taking the derivative of (6.3) with respect to α and setting it to zero yields

$$\log \alpha - \Psi(\alpha) = \log\left(\frac{\sum_{i=1}^{n_o} t_i}{n_o}\right) - \frac{1}{n_o}\left(\sum_{i=1}^{n_o} \log t_i\right), \tag{6.4}$$

where $\Psi(\alpha) = \dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the Digamma function. In order to estimate α, we need to solve (6.4) numerically, where the initial guess can be found using the method of moments. Choi & Wette (1969) found an explicit form for the Newton Raphson update of the initial guess, which is given by

$$\hat{\alpha}^{(k+1)} = \hat{\alpha}^{(k)} - \frac{\log \hat{\alpha}^{(k)} - \Psi\left(\hat{\alpha}^{(k)}\right) - s}{\left(\hat{\alpha}^{(k)}\right)^{-1} - \Psi'\left(\hat{\alpha}^{(k)}\right)}, \tag{6.5}$$

where s is the RHS of (6.4) and$\Psi'\left(\alpha^{(k)}\right)$ is the Trigamma function.

Employing the MLE, the estimates of α and θ are shown in Table 6.5. The KS test statistics are presented in Table 6.7, we can see that, for each passage, we cannot reject the null hypothesis that the incubation period data has a gamma distribution with the estimated parameters. Figure 6.6 shows the theoretical PDF plots of gamma distributions given by the estimates in Table 6.5. The estimates of α for each passage is larger than 1, so the distribution is bell-shaped for all passages.

Combined Data	$\hat{\alpha}$	$\hat{\theta}$	Mean	Mode	Variance
Passage 1	71.4294	0.1430	499.5063	492.5133	3493.05101
Passage 2	5.9826	0.0262	228.3435	190.1756	8715.4012
Passage 3	870.5195	5.3763	161.9180	161.7320	30.1170
Passage 4	132.5212	0.9346	141.7946	140.7246	151.7168
Passage 5	181.1971	1.2566	144.1963	143.4005	114.7512

Table 6.5: Fitting every passage's incubation period with a gamma distribution.

The incubation period from the first passage, $T_{g_1} \sim \Gamma(71.4294, 0.1430)$ is most probable at $T_{g_1} = 492$ days. It is not likely for the primary passage mice to show disease when T_{g_1} is less than 300 days. For the second passage, $T_{g_2} \sim \Gamma(5.9826, 0.0262)$ has a relatively small $\hat{\alpha}$ compared to the other passage. It means that the range of the incubation period is rather large. The estimate of α for the third passage's incubation period is the largest among all passages, we can see from Figure 6.6 that the range of T_{g_3} is small and the bell shape is peaked at $T_{g_3} = 161$ days. The probability of the fourth passage's incubation period is the highest when $T_{g_4} = 140$ days; while it is most likely for the fifth-passage mice to show disease after 143 days.

6.3.4 Weibull Model

We also find reasonable fit by applying a Weibull model to the incubation period of each passage. The Weibull distribution, with a positive rate parameter θ and a positive shape parameter α, has PDF

$$f(t;\theta,\alpha) = \alpha\theta^{\alpha}t^{\alpha-1}\exp\left[-(\theta t)^{\alpha}\right],$$

where $t \geq 0$, $\theta > 0$ and $\alpha > 0$. The Weibull distribution has been one of statisticians' favourite distributions for the fitting of lifetime data because it has a variety of different shapes, depending on the shape parameter α, and it provides an ease of fitting.

The estimates of the parameters can be obtained by maximising the following log-likelihood function:

$$l(\boldsymbol{\Theta}) = n_o \log(\alpha) + n\alpha \log(\theta) + \sum_{i=1}^{n_o}\left[(\alpha-1)\log(t_i) - (\theta t_i)^{\alpha}\right].$$

The MLE of α and θ are shown in Table 6.6; the mean, median mode and variance of the fitted Weibull distribution for each passage are shown in the same table. In Table 6.7, we show the KS tests results: for each passage, we cannot reject the null hypothesis that the incubation period comes from a Weibull distribution with the specified parameters. From Figure 6.7, we note that the shapes of the theoretical plots of the Weibull distributions are similar to the versions of the lognormal model, normal model and gamma model.

For first passage mice, the average incubation period is 498 days, while the mode is $T_{g_1} = 519$ days; The range (≈ 600 days) of the second passage's incubation period is the

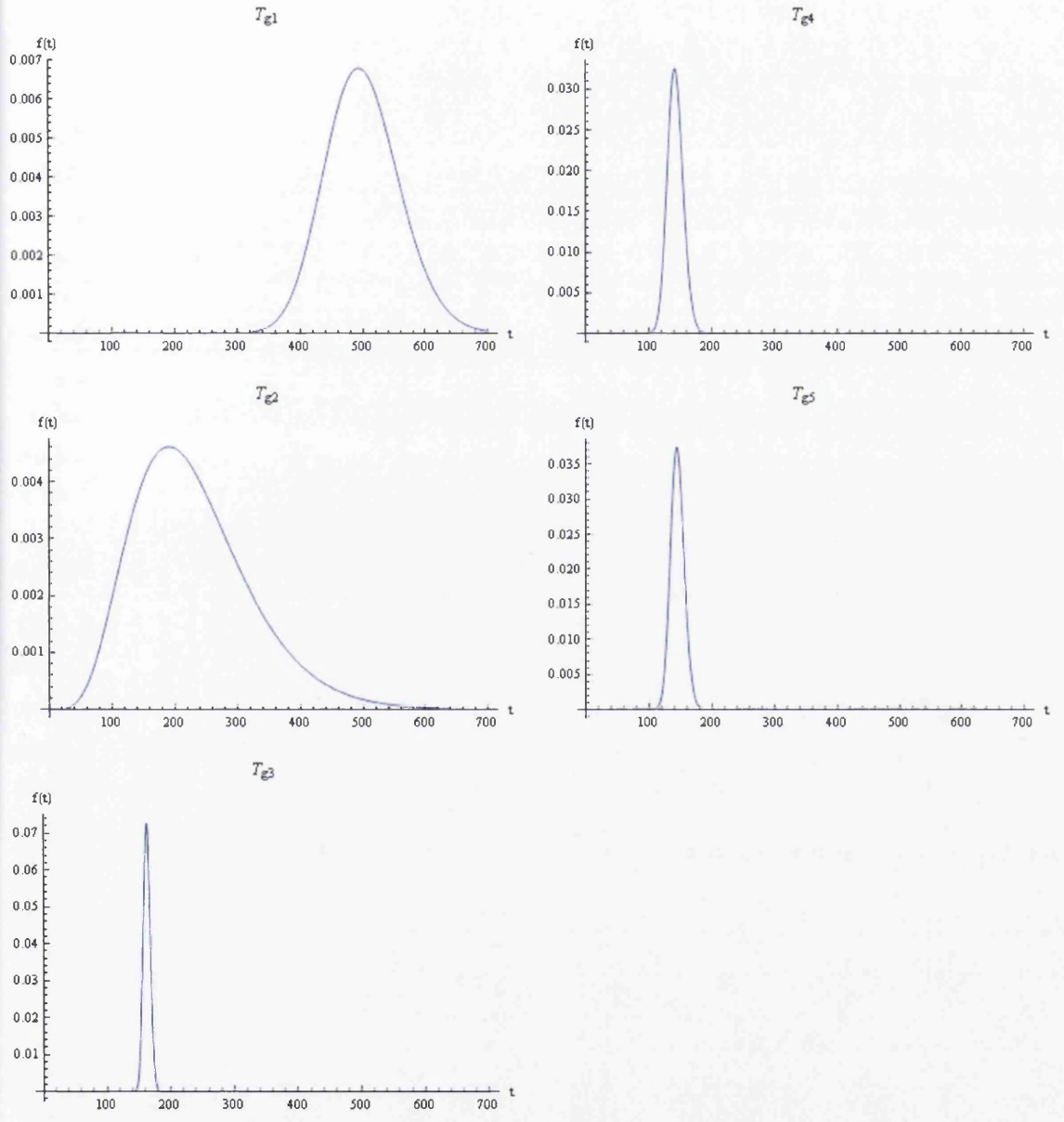


Figure 6.6: The gamma models for generation 1 to generation 5.

Combined Data	$\hat{\alpha}$	$\hat{\theta}$	Mean	Median	Mode	Variance
Passage 1	9.1081	0.0019	498.6716	505.5571	519.6380	4292.3011
Passage 2	2.4537	0.0039	227.4048	220.8331	207.1480	9791.1777
Passage 3	25.8968	0.0061	160.5140	161.6306	163.6853	59.9020
Passage 4	10.0750	0.0068	139.9487	141.8052	145.5409	279.5208
Passage 5	14.9669	0.0067	144.1189	145.6432	148.5657	139.5061

Table 6.6: Fitting every passage's incubation period with a Weibull distribution.

largest among all of the passages. On average, the incubation period of second passage mice is 227 days; whereas the mode is 207 days. The characteristics of the third, fourth and fifth generations' incubation periods are quite similar, where the average incubation periods are generally lower than 160 days; whereas the incubation period from these three passages are most probable when $T_{g_3} = 163$, $T_{g_4} = 145$ and $T_{g_5} = 148$ respectively. The incubation period from all passages have left-skewed distributions, except for the second passage. We compare Table 6.6 with Table 6.3 to 6.5, and find the variances suggested by the Weibull models are generally larger than the other models. The KS test results (see Table 6.7) of the Weibull models tell us that it is outperformed by the other models in terms of the goodness of fit.

6.3.5 Summary of Single Distribution Models

The estimation results for each distribution are summarised in Table 6.7. As seen in the table, all four models provide a reasonable fit to each passage. Compared to the other models, the lognormal model provides a better fit to all passages, except from the first passage, where the other three models have shorter KS distances. Surprisingly, the normal model returns competitive KS distance in each passage, except from the second passage; as we mentioned before, the fitted normal distribution of the second passage allows negative incubation periods, which are impossible in practice. This is why its significance value is the lowest ($sig = 0.3863$), less than half of the first passage's significance value. In general, the Weibull model has the largest KS distance for each passage (except from the second passage) compared to its rivals.

As a conclusion, all four models tell a similar story regarding the properties of each passage's incubation period. It takes a significantly longer time for first passage mice to show prion disease symptoms, compared to their successors. Second passage mice have a large range of incubation period and hence the variance of the incubation period is very much larger than the other passage's variance. The third, fourth and fifth passage mice have much lower incubation periods (on average between 140 to 160 days) and lower variation compared to their predecessor. The variance of the third passage incubation period is the smallest among all passages. In general, the incubation period of a mouse decreases with the number of the passage.

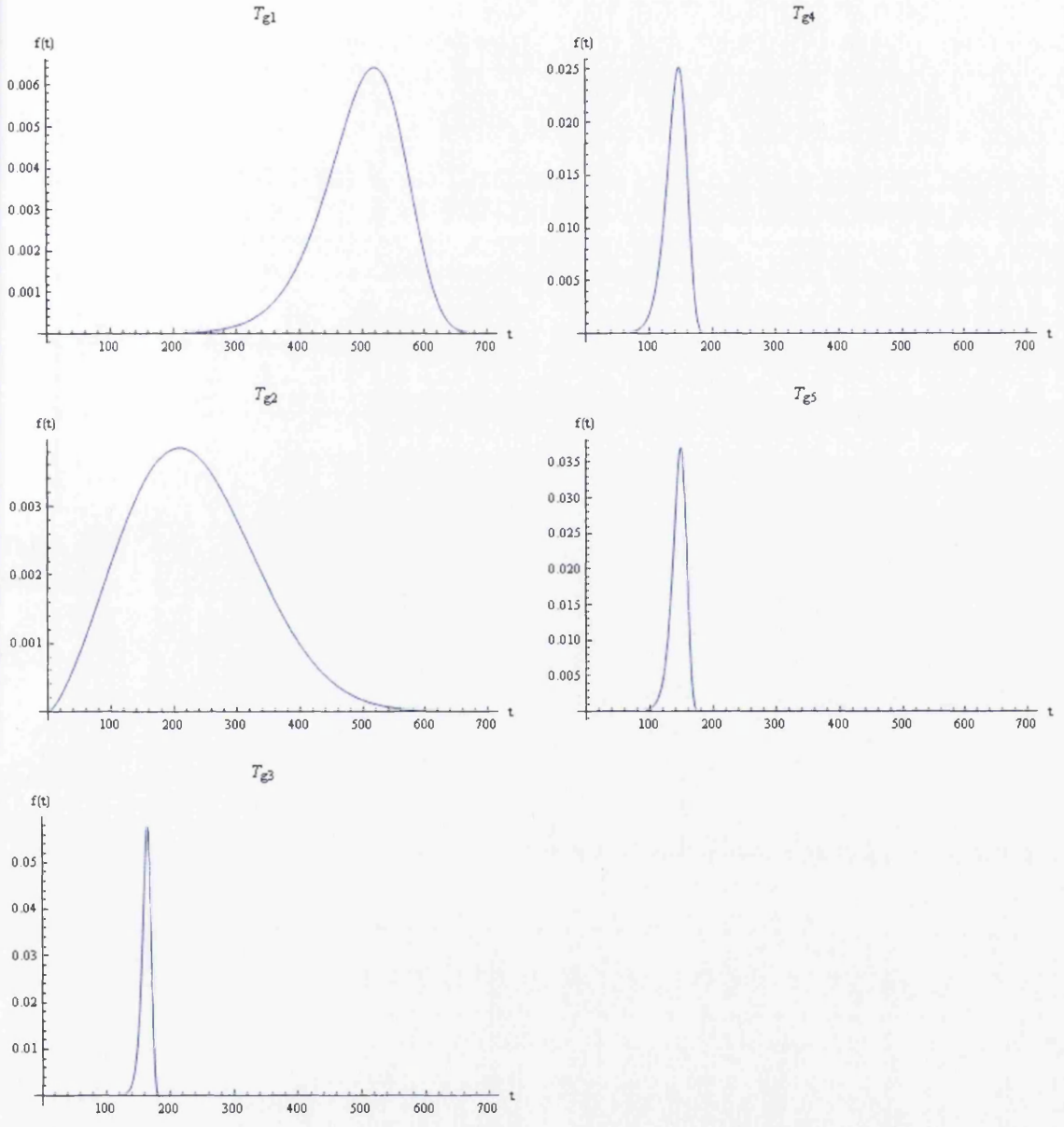


Figure 6.7: The Weibull models for generation 1 to generation 5.

Passage	Lognormal	Normal	Gamma	Weibull
1	$\hat{\mu} = 6.2068$ $\hat{\sigma} = 0.1251$ $KS = 0.2024$ $sig = 0.7366$	$\hat{\mu} = 499.6000$ $\hat{\sigma} = 62.2739$ $KS = 0.1826$ $sig = 0.8353$	$\hat{\alpha} = 71.4294$ $\hat{\theta} = 0.1430$ $KS = 0.1901$ $sig = 0.7989$	$\hat{\alpha} = 9.1081$ $\hat{\theta} = 0.0019$ $KS = 0.2012$ $sig = 0.7426$
2	$\hat{\mu} = 5.3448$ $\hat{\sigma} = 0.4182$ $KS = 0.1341$ $sig = 0.8999$	$\hat{\mu} = 228.3125$ $\hat{\sigma} = 102.9678$ $KS = 0.2164$ $sig = 0.3863$	$\hat{\alpha} = 5.9826$ $\hat{\theta} = 0.0262$ $KS = 0.1597$ $sig = 0.7521$	$\hat{\alpha} = 2.4537$ $\hat{\theta} = 0.0039$ $KS = 0.1962$ $sig = 0.5076$
3	$\hat{\mu} = 5.0862$ $\hat{\sigma} = 0.0360$ $KS = 0.3030$ $sig = 0.3782$	$\hat{\mu} = 161.8750$ $\hat{\sigma} = 5.9387$ $KS = 0.3086$ $sig = 0.3566$	$\hat{\alpha} = 870.5195$ $\hat{\theta} = 5.3763$ $KS = 0.3120$ $sig = 0.3437$	$\hat{\alpha} = 25.8968$ $\hat{\theta} = 0.0061$ $KS = 0.3284$ $sig = 0.2861$
4	$\hat{\mu} = 4.9506$ $\hat{\sigma} = 0.0954$ $KS = 0.3967$ $sig = 0.3177$	$\hat{\mu} = 141.8000$ $\hat{\sigma} = 14.3073$ $KS = 0.4047$ $sig = 0.2965$	$\hat{\alpha} = 132.5212$ $\hat{\theta} = 0.9346$ $KS = 0.4110$ $sig = 0.2806$	$\hat{\alpha} = 10.0750$ $\hat{\theta} = 0.0068$ $KS = 0.4128$ $sig = 0.2762$
5	$\hat{\mu} = 4.9684$ $\hat{\sigma} = 0.0829$ $KS = 0.2902$ $sig = 0.7024$	$\hat{\mu} = 144.2000$ $\hat{\sigma} = 12.0291$ $KS = 0.2969$ $sig = 0.6763$	$\hat{\alpha} = 181.1971$ $\hat{\theta} = 1.2566$ $KS = 0.3130$ $sig = 0.6127$	$\hat{\alpha} = 14.9669$ $\hat{\theta} = 0.0067$ $KS = 0.3351$ $sig = 0.5261$

Table 6.7: Comparison of the performances of all single distribution models fitted to each passage's incubation period.

6.4 Fitting the Incubation Period Data with Mixture Distributions

A number of different distributions have been tried to fit the overall incubation period data, however, none of the single distributions appear to provide a significant fit to the incubation period data. For example, from previous section, we learned that the first passage incubation periods are significantly different from the others. This makes us think that the overall data might come from a mixture of two distributions. In the following subsections, we show that a few mixture models can be used to explain important features of the incubation period.

6.4.1 Mixtures of Lognormal Distributions

The single lognormal distribution has long been used in modelling incubation period data. First, we combined the incubation period from different passages into one sample of size $n_o = 44$, and fit the data with a single lognormal distribution; unfortunately, the goodness of fit of such a fitted distribution is poor. The parameter estimates of a single lognormal distribution are $\hat{\mu} = 5.4061$ and $\hat{\sigma} = 0.5310$; where the KS test showed that the departure of the ECDF from the theoretical CDF of such a lognormal distribution ($KS = 0.2473$) is significant ($sig = 0.0074$). Therefore, we fit the incubation period with a mixture of two

lognormal distributions, using the MLE. The log-likelihood function of a mixture of two lognormal distributions, with the parameter space $\Theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, p)$, is

$$l(\Theta) = \sum_{i=1}^{n_o} \log \left[\frac{p}{t_i \sigma_1 \sqrt{2\pi}} \exp \left[-\frac{(\log[t_i] - \mu_1)^2}{2\sigma_1^2} \right] + \frac{(1-p)}{t_i \sigma_2 \sqrt{2\pi}} \exp \left[-\frac{(\log[t_i] - \mu_2)^2}{2\sigma_2^2} \right] \right] \tag{6.6}$$

Using Mathematica, we find the estimates of Θ which maximise (6.6) numerically, and the estimation results are as follows:

$$\begin{aligned} \hat{\mu}_1 &= 5.0108, \\ \hat{\sigma}_1 &= 0.1070, \\ \hat{\mu}_2 &= 5.9624, \\ \hat{\sigma}_2 &= 0.3430, \\ \hat{p} &= 0.5846. \end{aligned} \tag{6.7}$$

The maximum log-likelihood given by the estimates in (6.7) is

$$l\left(\hat{\Theta}\right) = -251.304. \tag{6.8}$$

The KS plot of this lognormal mixture model is shown in Figure 6.8; the KS test shows that the mixture lognormal model with the parameters in (6.7) has a KS distance 0.1223 and $sig = 0.4991$. In other words, a mixture of two lognormal distribution provides a reasonable fit to the full set of incubation period data across all passages. The plots (in Figure 6.9), which compare the theoretical PDF of a mixture of two lognormal distributions with the parameters in (6.7) and the histogram of the incubation period, further confirm the reasonable fit of the lognormal mixture model.

According to the lognormal mixture model, the first component of the mixture distribution has a relatively shorter mean, 150 days and a relatively low variance, 262; whereas the second component has a larger mean, 412 days and a variance, 21205. The probability that the incubation period is drawn from the first component is a 0.5846.

Next, we assume that the incubation period of each passage has a mixture of two lognormal distribution with the parameter values stated in (6.7) and an unknown mixing weight p, which is then estimated numerically using MLE. The estimates of p for each passage are shown in Table 6.8 and Figure 6.10. The results show that the incubation period of passage 1 has $\hat{p} = 0$, indicating that it derives from the second component only; passage 2 has a $\hat{p} = 0.4865$, with a low KS distance, this means that the incubation period of passage 2 derives from a roughly equal mixture of two lognormal distributions. The incubation periods of the 3^{rd}, 4^{th} and 5^{th} passage have $\hat{p} = 1$, indicating that they are drawn from the first component only. We now have a brief picture that the first passage's incubation period is likely to come from the second component of the lognormal mixture model, where the

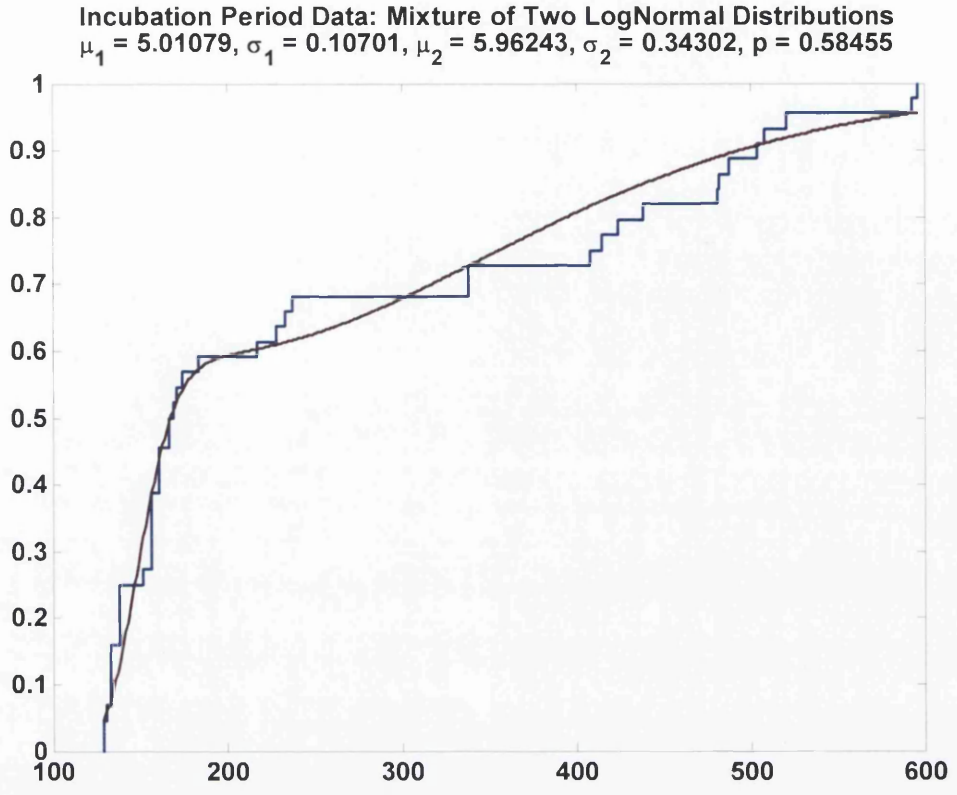


Figure 6.8: Comparison of the ECDF plot of the incubation period data and the fitted CDF plot given by a two-component lognormal mixture model.

incubation period is relatively longer on average with a high variance; the second passage's incubation period is fitted well with a mixture of two lognormal distributions; whereas the incubation period of the generation after the second passage is likely to be drawn from the first component of the lognormal mixture model, in which the average incubation period is relatively shorter and the variance is also relatively smaller.

6.4.2 Mixtures of Normal Distributions

By now, we have learned that the incubation period data can be well described by a mixture of two lognormal distributions. In previous sections, we have seen that there are other distributions which can be used to fit the incubation period from each passage. Therefore, we should consider other mixture models for the joint data in order to find the best fitting for the incubation period data. In this section, we fit the data with a mixture of two normal distributions. The goodness of fit of such a normal mixture model appears to be satisfactory. The parameters $\Theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, p)$ which maximise the likelihood function of a mixture

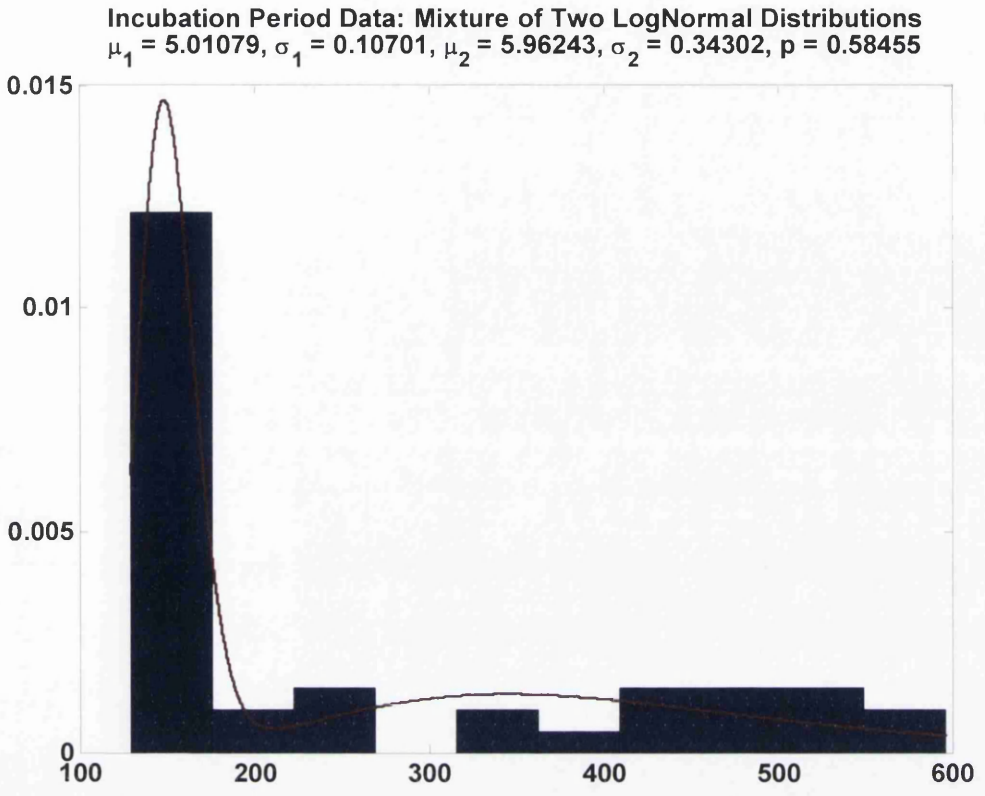


Figure 6.9: Comparison of the EPDF plot of the incubation period data and the fitted PDF plot given by a two-component lognormal mixture model.

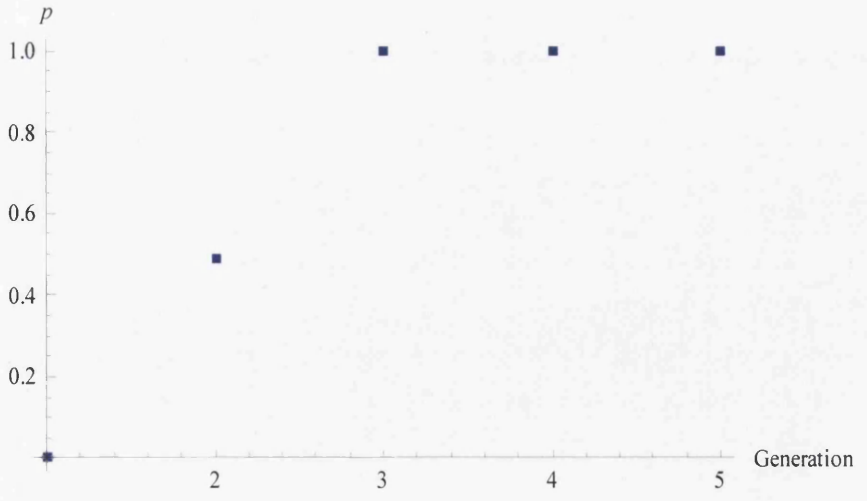


Figure 6.10: Fitting a mixture of two lognormal distributions to every generation's incubation period: Plot of $\hat{p}$ versus Generation.

Passage	$\hat{p}$	KS	sig
1	0	0.5566	0.0019
2	0.4865	0.2251	0.3399
3	1	0.6645	0.0005
4	1	0.5825	0.0385
5	1	0.3825	0.3589

Table 6.8: Fitting a mixture of two lognormal distributions to every generation with p unknown.

of two normal distributions

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{n_o} \log \left[p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left[\frac{t_i - \mu_1}{\sigma_1}\right]^2 \right) + (1-p)\frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left[\frac{t_i - \mu_2}{\sigma_2}\right]^2 \right) \right] \tag{6.9}$$

are

$$\begin{aligned} \hat{\mu}_1 &= 150.74, \\ \hat{\sigma}_1 &= 15.889, \\ \hat{\mu}_2 &= 405.29, \\ \hat{\sigma}_2 &= 126.31, \\ \hat{p} &= 0.5767. \end{aligned} \tag{6.10}$$

The maximum log-likelihood given by the estimates in (6.10) is

$$l\left(\hat{\boldsymbol{\Theta}}\right) = -249.912. \tag{6.11}$$

From (6.10) we can see that, according to the normal mixture model, the distribution of the incubation period data has two normal components, where the first component has a lower mean period ($\approx$ 151 days) and a smaller variance while the second component has a longer average incubation period ($\approx$ 405 days) with a much higher variance. The probability that an observation is drawn from the first component is 0.5767, suggesting that more observations come from the first component which has a shorter incubation period. In order to assess the goodness of fit, we produce the KS plot, as shown in Figure 6.11, the maximum distance between the empirical and fitted cumulative probability distribution plots is $KS = 0.1209$, the test statistic of this model is smaller than the mixture lognormal model ($KS = 0.1223$). The asymptotic significance value is 0.5145. From Figure 6.12, we can see that the theoretical PDF plot fits quite well to the histogram.

Now, we assume that each passage has a mixture of two normal distribution where the parameters are known as the ones in (6.10) except from p. With only one unknown parameter to be estimated, we calculate p which maximises the likelihood function as in (6.9) for each passage. The estimates of p for each passage are shown in Table 6.9 and

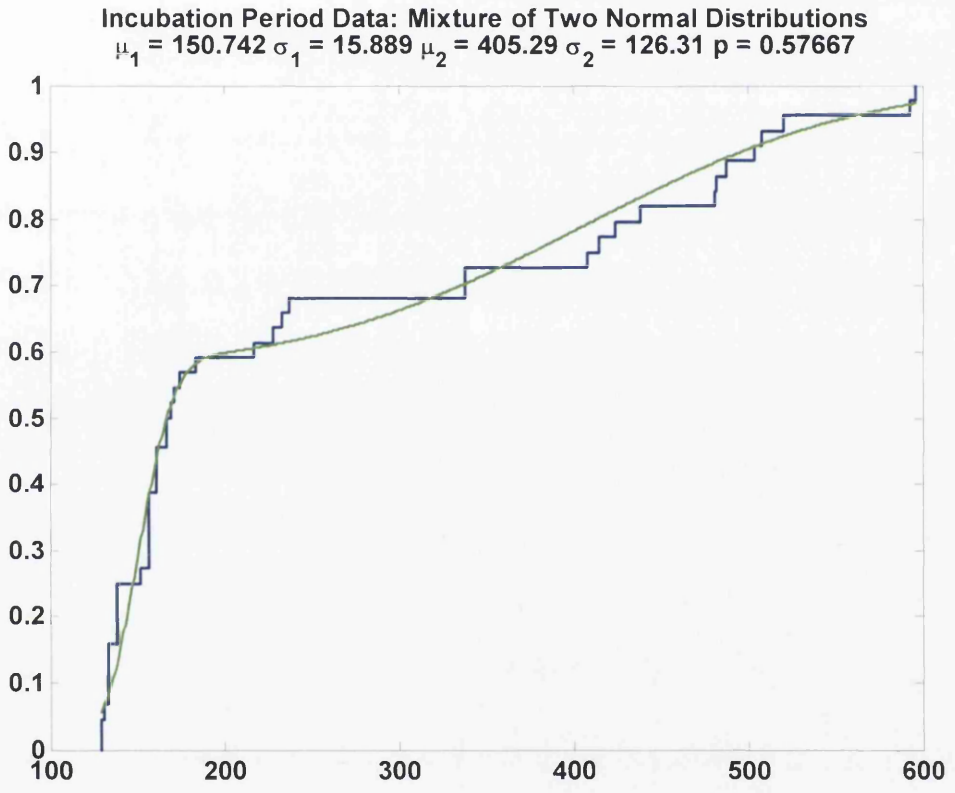


Figure 6.11: Comparison of the ECDF plot of the incubation period data and the fitted CDF plot given by a two-component normal mixture model.

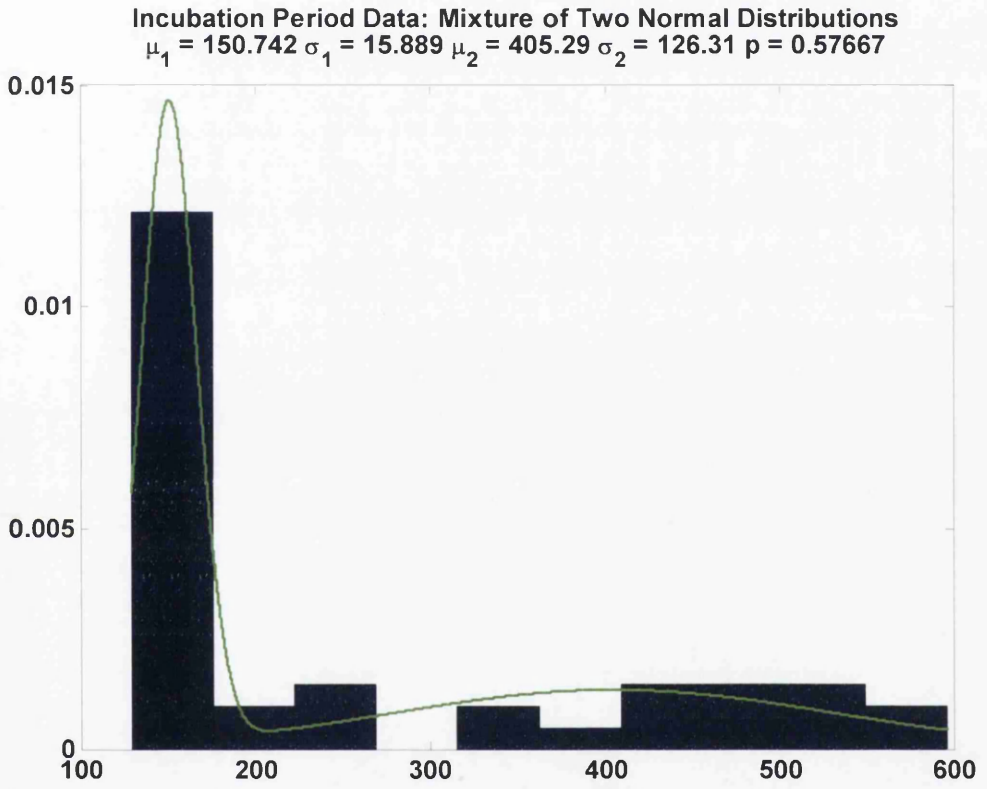


Figure 6.12: Comparison of the EPDF plot of the incubation period data and the fitted PDF plot given by a two-component normal mixture model.

Passage	$\hat{p}$	KS	sig
1	0	0.5255	0.0042
2	0.4725	0.2458	0.2451
3	1	0.6532	0.0007
4	1	0.5887	0.0353
5	1	0.3887	0.3402

Table 6.9: Fitting a mixture of two normal distributions to every generation with p unknown.

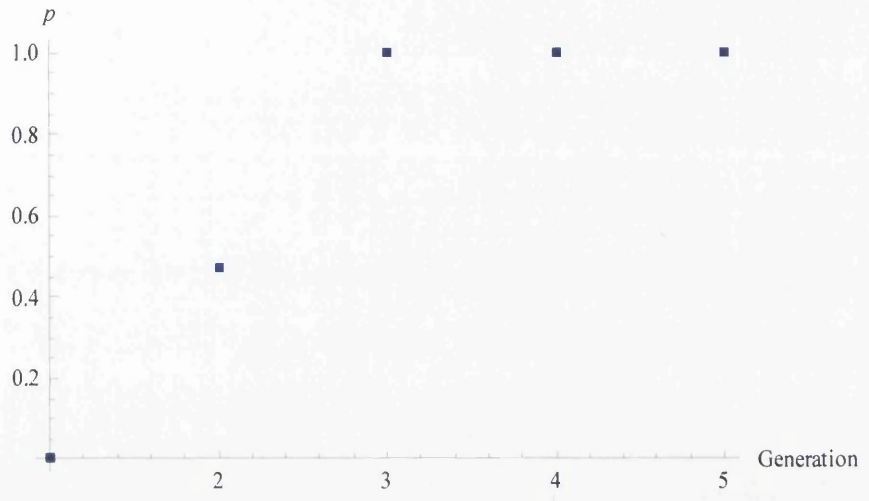


Figure 6.13: Fitting a mixture of two normal distributions to every generation's incubation period: Plot of $\hat{p}$ versus Generation.

Figure 6.13. The results are similar to the lognormal mixture version in Table 6.8.

The normal mixture model appears to have a marginally better fit than the lognormal mixture model (the significance of the former is only slightly higher than the latter). The normal mixture distribution also has a slightly larger likelihood than the lognormal mixture(compare (6.8) to (6.11)). Both of the mixture models suggest that the underlying distribution of the incubation period is a mixture of two distributions. Both models agree that the first component has a shorter average incubation period with a smaller variation while the other component has a longer incubation period with a larger variation. Both give similar estimate of the mixing weight at first passage, $\hat{p} \approx 0.58$.

6.4.3 Mixtures of Gamma Distributions

In this section, we fit the overall data set with a mixture of two gamma distributions. Like before, we denote by p the prior probability that an incubation period is drawn from the first gamma distribution. We now have $\mathbf{\Theta} = (\alpha_1, a, \alpha_2, b, p)$, where α_1 and a are the shape parameter and the scale parameter of the first density component respectively; α_2 and b are the parameters of the second density component. Estimating the parameters of mixtures of gamma distributions is not straightforward, and there are five parameters to be estimated

from a mixture with two gamma components. Using Mathematica, we numerically maximise the log-likelihood function, $l(\boldsymbol{\Theta})$, where

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{n_o} \log\left[p\frac{a^{\alpha_1}}{\Gamma(\alpha_1)}\exp(-at_i)\,t_i^{\alpha_1-1} + (1-p)\frac{b^{\alpha_2}}{\Gamma(\alpha_2)}\exp(-bt_i)\,t_i^{\alpha_2-1}\right].$$

The estimates of the parameters are

$$\begin{aligned} \hat{\alpha}_1 &= 88.998, \\ \hat{a} &= 0.5903, \\ \hat{\alpha}_2 &= 9.0820, \\ \hat{b} &= 0.0223, \\ \hat{p} &= 0.5805. \end{aligned} \tag{6.12}$$

The maximum log-likelihood given by the estimates in (6.12) is

$$l\left(\hat{\boldsymbol{\Theta}}\right) = -250.627. \tag{6.13}$$

The KS test shows a reasonable goodness of fit to the incubation data given by the estimates in (6.12). The maximum distance between the ECDF and the theoretical CDF is 0.1231 (the KS plots are shown in Figure 6.14) and the asymptotic significance value is $sig = 0.4910$. The theoretical PDF given by (6.12) is plotted along with the histogram of the observed data in Figure 6.15; it is clear that the mixture model again fits nicely to the observed data.

Once again, it seems like there exists two types of characteristics, one is gamma distributed with $\hat{\alpha}_1 = 88.9980$, $\hat{a} = 0.5903$ and the other one is gamma distributed with $\hat{\alpha}_2 = 9.0820$, $\hat{b} = 0.0223$. The probability of an incubation period to be drawn from the first gamma component is $\hat{p} = 0.5805$. The gamma mixture model suggests that when the incubation period T is longer, the shape parameter α is smaller and hence the variance is larger. Conversely, for smaller T, α is larger with smaller variance of T. If we assume that each passage has a mixture of two gamma distribution with unknown p, we only need to estimate the proportion of the first gamma component for every passage. The estimates of p for each passage are shown in Table 6.10 and Figure 6.16. Passage 1 has $\hat{p} = 0$, indicating that it comes from the second component; passage 2 can be fitted well with this mixture gamma model with $\hat{p} = 0.4793$; passage 3, 4 and 5 are drawn from the first component as all of them have $\hat{p} = 1$. However, the KS test statistics are significantly large for passage 1 and 3.

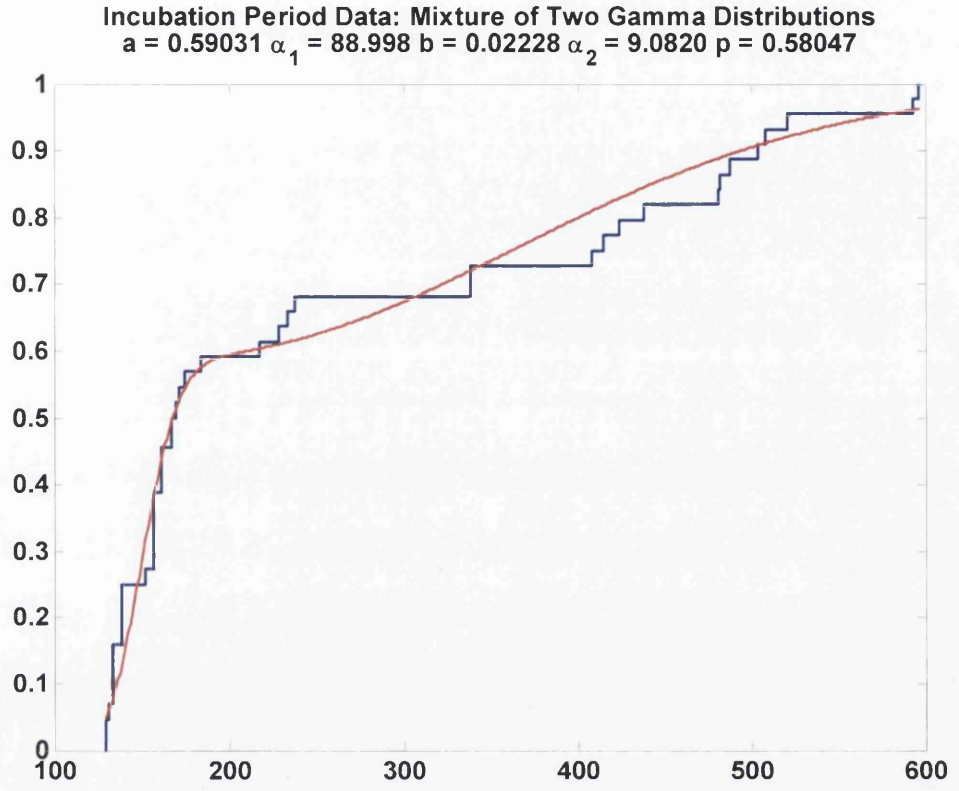


Figure 6.14: Comparison of the ECDF plot of the incubation period data and the fitted CDF plot given by a two-component gamma mixture model.

Passage	$\hat{p}$	KS	sig
1	0	0.5454	0.0026
2	0.4793	0.2274	0.3283
3	1	0.6628	0.0006
4	1	0.5843	0.0375
5	1	0.3843	0.3533

Table 6.10: Fitting a mixture of two gamma distributions to every generation with p unknown.

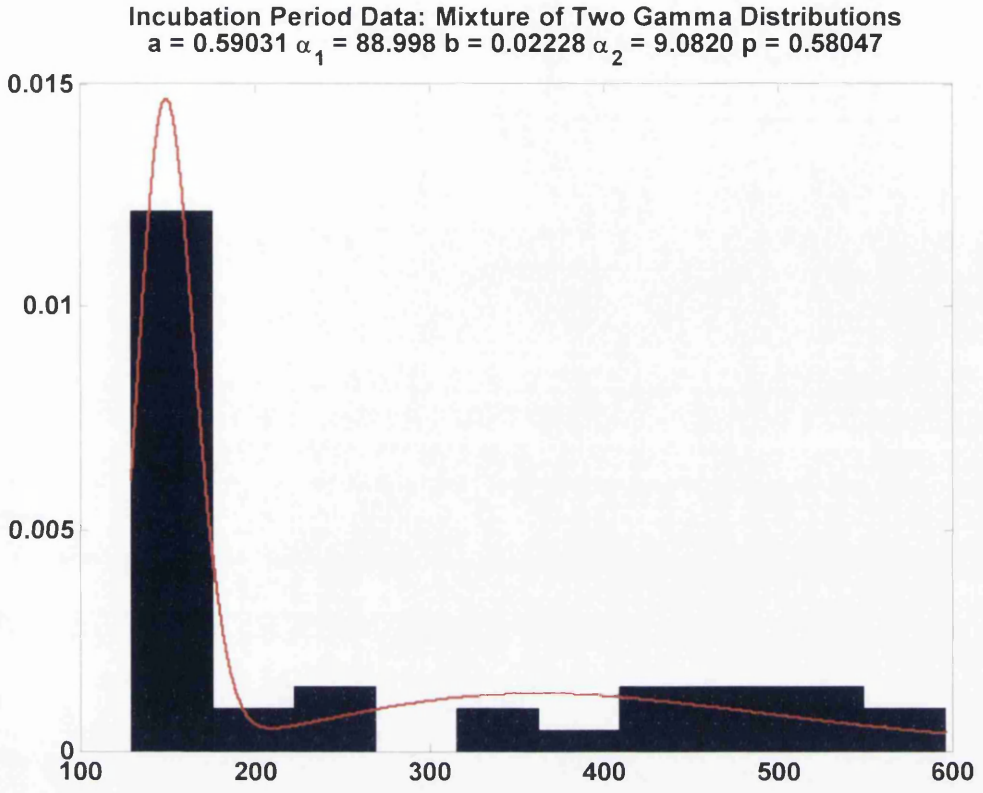


Figure 6.15: Comparison of the ECDF plot of the incubation period data and the fitted CDF plot given by a two-component gamma mixture model.

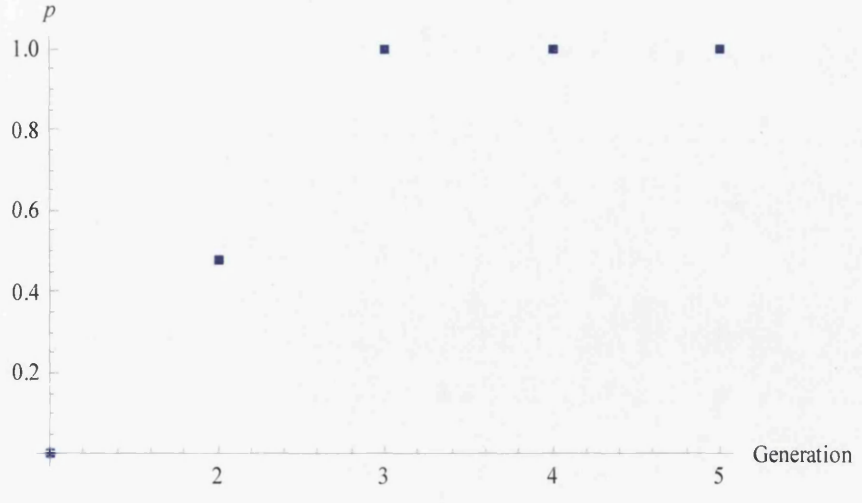


Figure 6.16: Fitting a mixture of two gamma distributions to every generation's incubation period: Plot of $\hat{p}$ versus Generation.

6.4.4 Mixtures of Weibull Distributions

We investigated the fitting of Weibull mixture model to the incubation period data and found good agreement between the model and the observed data. An observation from a mixture of two Weibull distributions has density

$$f(t;\Theta) = p\alpha_1 a^{\alpha_1} t^{\alpha_1 - 1} \exp\left[-(at)^{\alpha_1}\right] + (1-p)\,\alpha_2 b^{\alpha_2} t^{\alpha_2 - 1} \exp\left[-(bt)^{\alpha_2}\right]$$

where $\Theta = (a, b, \alpha_1, \alpha_2, p)$, a and b are the distinct rate parameters of the two components, α_1 and α_2 are the shape parameters, and p is the mixing weight of the first Weibull component. Considering the log-likelihood function using the incubation period data, we have

$$l(\Theta) = \sum_{i=1}^{n_o} \log\left[p\alpha_1 a^{\alpha_1} t_i^{\alpha_1 - 1} \exp\left[-(at_i)^{\alpha_1}\right] + (1-p)\,\alpha_2 b^{\alpha_2} t_i^{\alpha_2 - 1} \exp\left[-(bt_i)^{\alpha_2}\right]\right]. \quad (6.14)$$

By maximising (6.14) numerically, we obtained the estimates of the parameters as follows

$$\begin{aligned} \hat{a} &= 0.0063, \\ \hat{\alpha}_1 &= 10.521, \\ \hat{b} &= 0.0022, \\ \hat{\alpha}_2 &= 3.7656, \\ \hat{p} &= 0.5758, \end{aligned} \quad (6.15)$$

while the log-likelihood given by the estimates in (6.15) is

$$l\left(\hat{\Theta}\right) = -250.18. \quad (6.16)$$

The Weibull mixture model suggests that 58% of the exposed hosts have a shorter average incubation period of 151 days, whereas 42% of the infected hosts live longer with a mean incubation period of 407 days. Similarly to other mixture models, the first component has a lower variation than the second component, as seen in Figure 6.18, which compares the theoretical PDF plot with the histogram of the observed data. The theoretical CDF plot using the estimates shown in (6.15) is presented in Figure 6.17 together with ECDF plot of the incubation period data. The KS test was taken to assess the goodness of fit of the Weibull mixture model, the test statistic KS is 0.1210, yielding a significance value of 0.5132. Comparing the significance value to the one given by the normal mixture model, we found that the Weibull version has a similar goodness of fit as the normal version. Since both models fit equally well to the incubation period data, the Weibull mixture model is preferred because the incubation period is not allowed to be negative.

Now, we assume that each passage has a mixture of two Weibull distributions where the parameters are known as the ones in (6.15) except for p. With only one unknown parameter

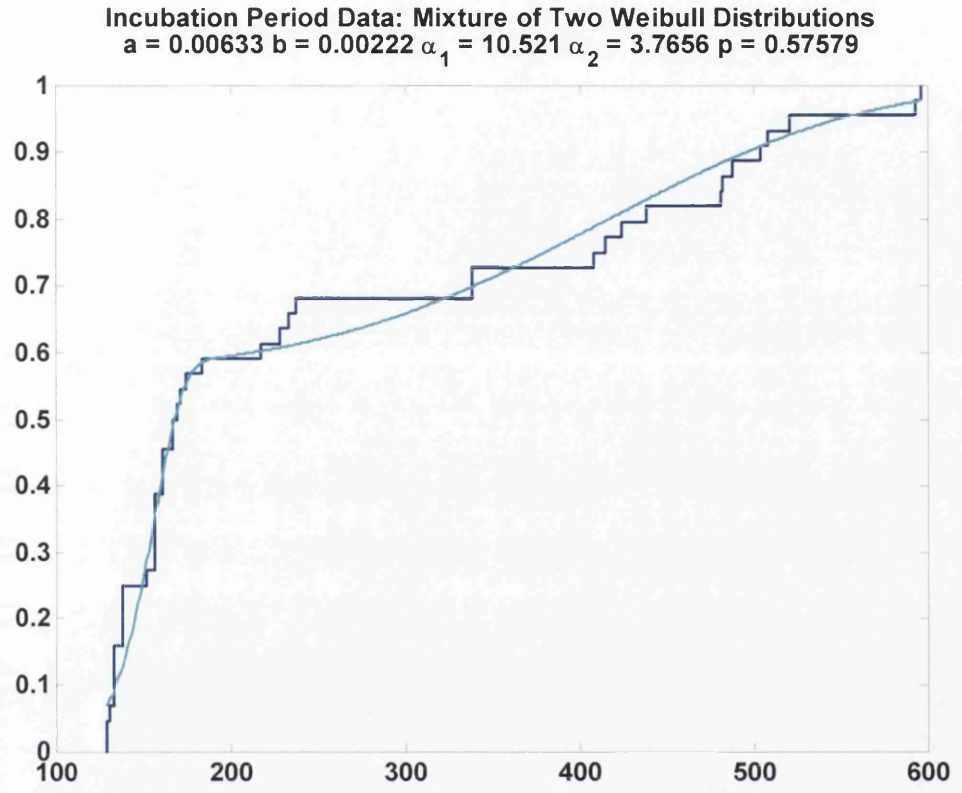


Figure 6.17: Comparison of the ECDF plot of the incubation period data and the fitted CDF plot given by a two-component Weibull mixture model.

Passage	$\hat{p}$	KS	sig
1	0	0.5240	0.0044
2	0.4681	0.1994	0.4874
3	1	0.6104	0.0022
4	1	0.5845	0.0374
5	1	0.3896	0.3377

Table 6.11: Fitting a mixture of two Weibull distributions to every generation with p unknown.

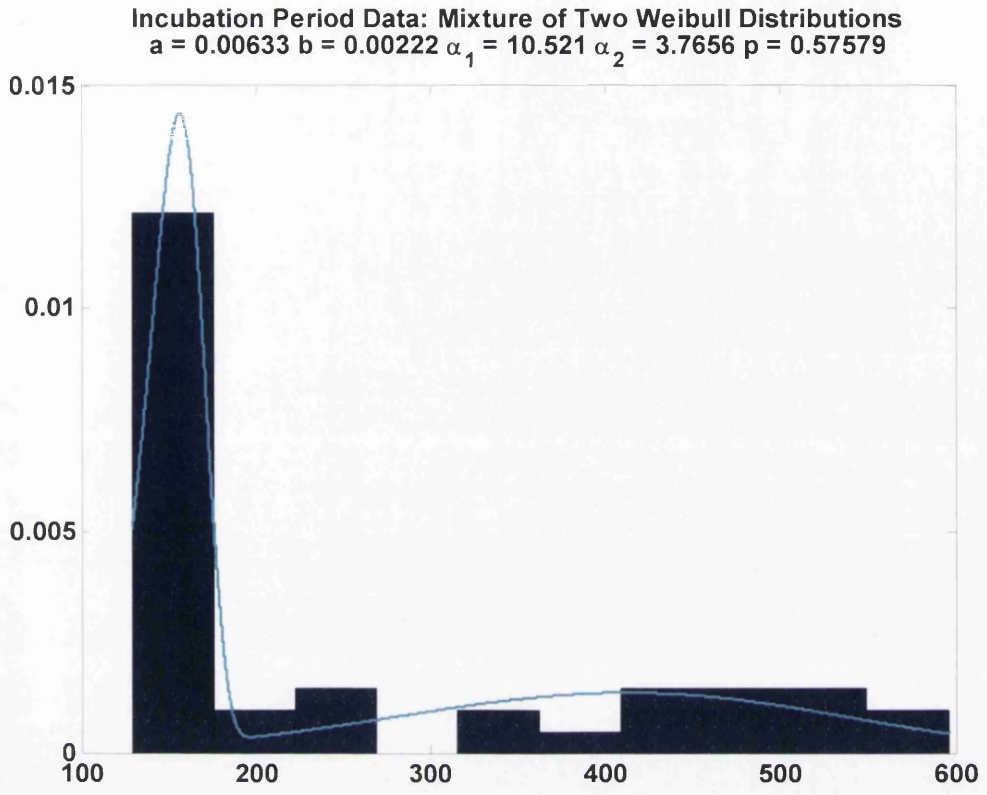


Figure 6.18: Comparison of the EPDF plot of the incubation period data and the fitted PDF plot given by a two-component Weibull mixture model.

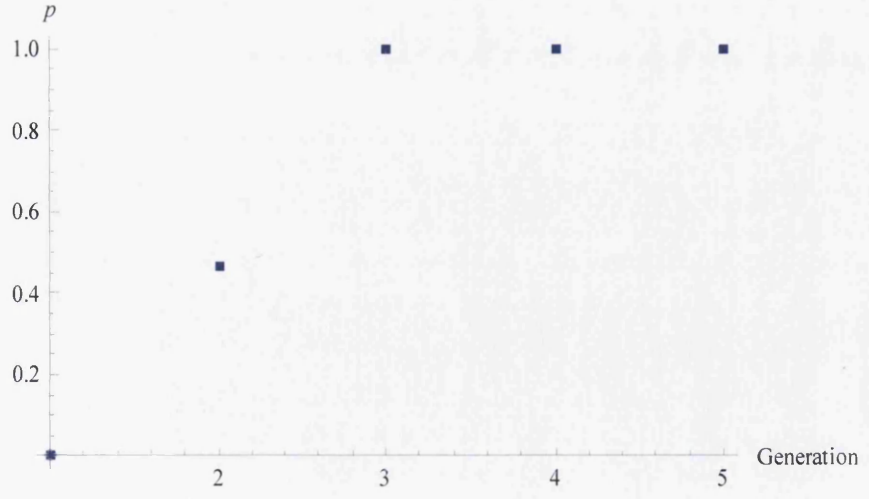


Figure 6.19: Fitting a mixture of two Weibull distributions to every generation's incubation period: Plot of $\hat{p}$ versus Generation.

to be estimated, we calculate p which maximises the likelihood function as in (6.14) for each passage. The estimates of p for each passage are shown in Table 6.11. From this table, we observe that the first passage has an incubation period following the second component of the Weibull mixture model. The incubation periods of the second passage has a mixture of two Weibull distributions with $\hat{p} = 0.4681$; the third, fourth and fifth passage's incubation periods are distributed according to the first component of the Weibull mixture model only.

6.4.5 Mixtures of Burr XII Distributions

The three parameter Burr XII distribution has become increasingly popular in the analysis of lifetime data. The Weibull distribution is a special case of the Burr XII distribution (see Watkins (1999)), with the following relationship

$$F_B\left(t;\alpha,\tau,\theta\right) = F_W\left(t;\tau,\theta^{-1}\alpha^{-\frac{1}{\tau}}\right), \tag{6.17}$$

where F_B denotes the CDF of a Burr XII distribution and F_W represents the CDF of a Weibull distribution. The Weibull distribution with shape parameter τ and scale parameter $\theta^{-1}\alpha^{-\frac{1}{\tau}}$ emerges as the limiting distribution of the Burr XII distribution. Given such a relationship, we speculate that a Burr XII mixture model, like its special case Weibull distribution, will also have a reasonable fit to the incubation period data. We now define a two-component Burr XII mixture model with density function

$$f\left(t;\Theta\right) = p\frac{\tau_1\alpha_1 a^{\tau_1}t^{\tau_1-1}}{\left(1+(at)^{\tau_1}\right)^{\alpha_1+1}} + (1-p)\frac{\tau_2\alpha_2 b^{\tau_2}t^{\tau_2-1}}{\left(1+(bt)^{\tau_2}\right)^{\alpha_2+1}},$$

where the parameter vector is $\Theta = (\alpha_1, \tau_1, a, \alpha_2, \tau_2, b, p)$. Such a mixture model requires us to estimate seven parameters with the 44 available observed data. We first use the optimisation tool in Mathematica to find the MLE $\hat{\Theta}$ which maximise the following log-likelihood function

$$l\left(\Theta\right) = \sum_{i=1}^{n_o}\log\left[p\frac{\tau_1\alpha_1 a^{\tau_1}t_i^{\tau_1-1}}{\left(1+(at_i)^{\tau_1}\right)^{\alpha_1+1}} + (1-p)\frac{\tau_2\alpha_2 b^{\tau_2}t_i^{\tau_2-1}}{\left(1+(bt_i)^{\tau_2}\right)^{\alpha_2+1}}\right]. \tag{6.18}$$

The algorithm does not converge to the tolerance level in 500 iterations, so we take the best estimated solution with feasibility residuals returned by Mathematica. The crude estimates

of Θ are as follows:

$$\begin{aligned} \hat{\alpha}_1 &= 918.48, \\ \hat{\tau}_1 &= 10.498, \\ \hat{a} &= 0.0033, \\ \hat{\alpha}_2 &= 492.21, \\ \hat{\tau}_2 &= 3.7686, \\ \hat{b} &= 0.0004, \\ \hat{p} &= 0.5759, \end{aligned} \tag{6.19}$$

while the log-likelihood given by the estimates in (6.19) is

$$l\left(\hat{\Theta}\right) = -250.18. \tag{6.20}$$

In order to compare this model to other mixture models studied earlier, we calculate the KS distance using the estimates in (6.19). The KS test statistic is 0.1205 and the significance value is 0.5188. The KS plot is shown in Figure 6.20; whereas the theoretical PDF plot is shown in Figure 6.21 together with the histogram of the observed data. The Burr XII mixture model is obviously telling a similar story like other mixture models considered in previous sections. In fact, it has the smallest KS distance among all mixture models considered here, in spite of the convergence difficulties in finding the MLEs.

Now, we assume that each passage has a mixture of two Burr XII distributions where the parameters are known as the ones in (6.19) except for p. With only one unknown parameter to be estimated, we calculate p which maximises the likelihood function as in (6.18) for each passage. The estimates of p for each passage are shown in Table 6.12 and Figure 6.22. Similarly to the other mixture models, the results show that the first passage's incubation period comes from the second component; the second passage's incubation period can be represented by a mixture of two Burr XII distributions; while the final three passages have incubation periods which follow the first component of the Burr XII mixture model. It should be noted that the second passage incubation period is marginally better described by the Burr XII mixture model than the mixture models discussed in previous sections, as indicated by a shorter KS distance and a larger significance value (compare Tables 6.12 to 6.11).

indicated by a shorter KS distance and a larger significance value (compare Tables 6.12 to

6.5 Summary

The combined incubation period data appears to be well fitted by all the mixture models discussed here. We now compare the models with the purpose of choosing the best model. Table 6.13 summarises the performance of each model. The estimated parameters, the KS test statistics and its significance value of each model are shown in the table. The estimated

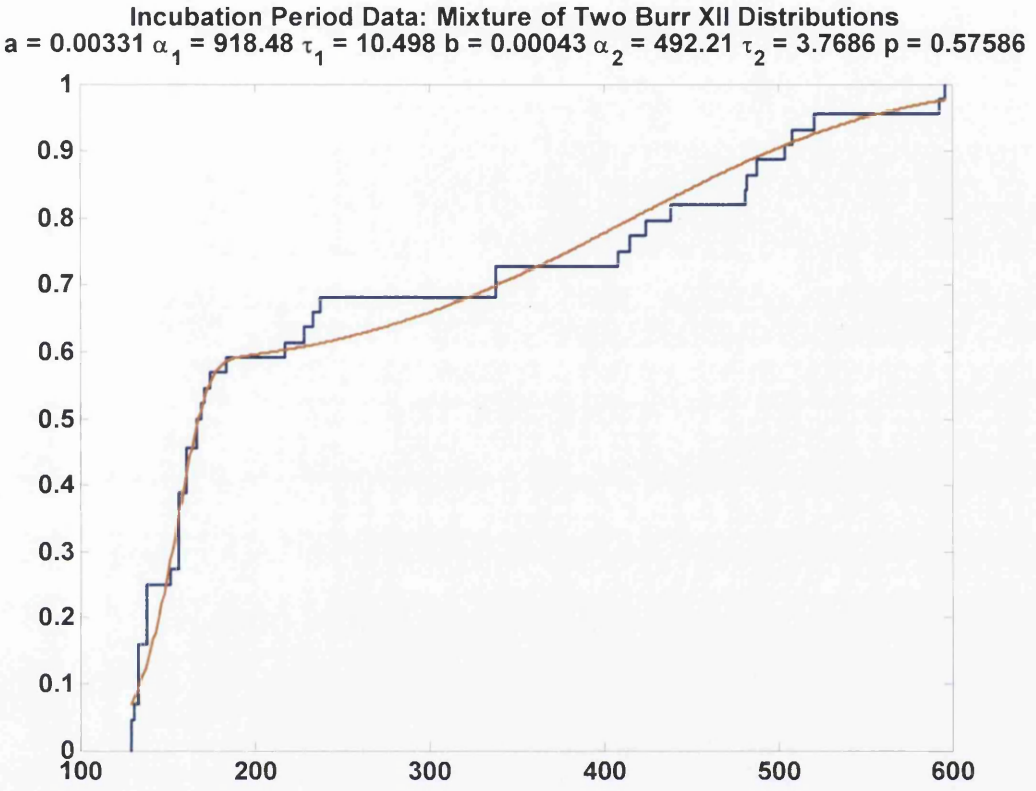


Figure 6.20: Comparison of the ECDF plot of the incubation period data and the fitted CDF plot given by a two-component Burr XII mixture model.

Passage	$\hat{p}$	KS	sig
1	0	0.5240	0.0044
2	0.4683	0.1993	0.4879
3	1	0.6109	0.0021
4	1	0.6170	0.0099
5	1	0.3392	0.3891

Table 6.12: Fitting a mixture of two Burr XII distributions to every generation with p unknown.

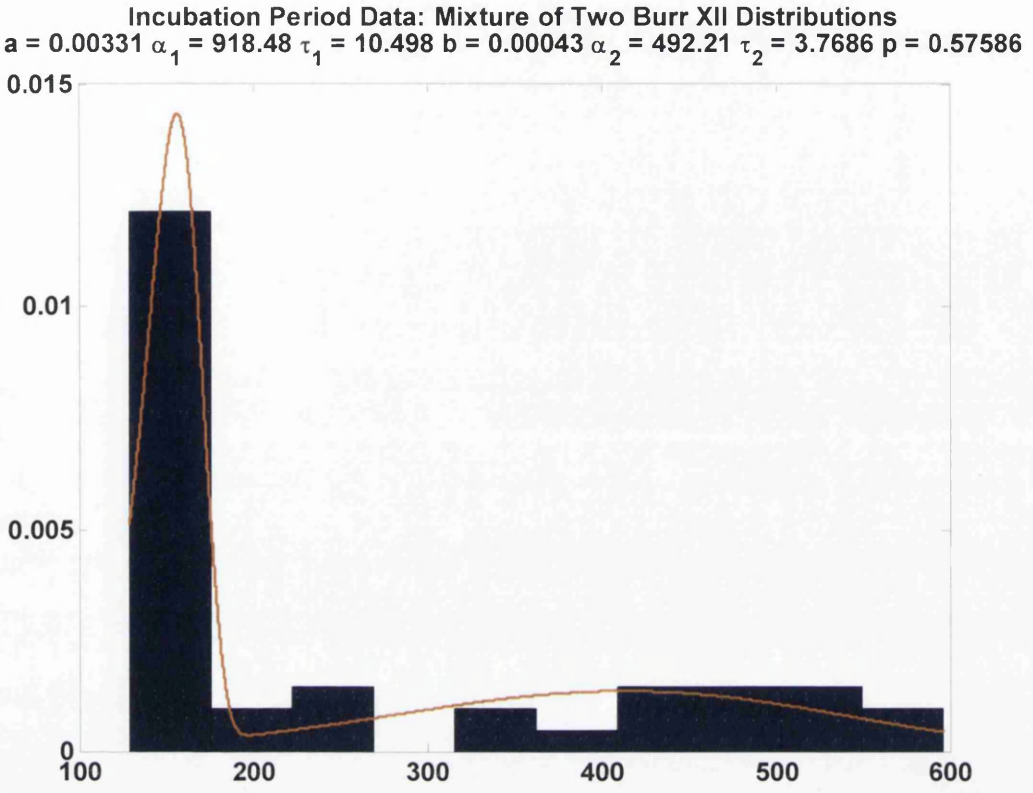


Figure 6.21: Comparison of the ECDF plot of the incubation period data and the fitted CDF plot given by a two-component Burr XII mixture model.

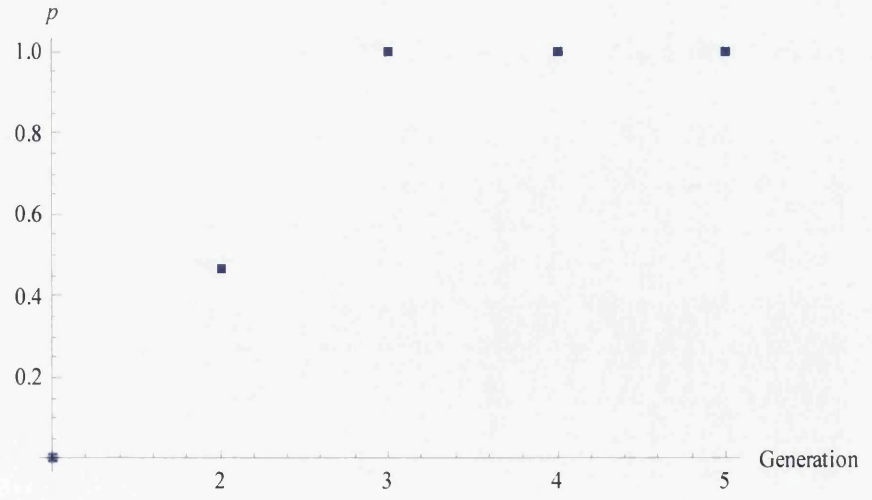


Figure 6.22: Fitting a mixture of two Burr XII distributions to every generation's incubation period: Plot of $\hat{p}$ versus Generation.

mean and variance of the incubation period for each component are also shown. The gamma mixture model appears to be slightly inferior to other models because it has the largest KS distance; whereas the Burr XII model outperforms other mixture models by giving the lowest KS distance. The models are also compared visually using the KS plots in Figure 6.23 and the PDF plots in Figure 6.24. We can see that for longer incubation periods, the lognormal mixture model has a good agreement with the gamma mixture model; whereas the normal mixture model, Weibull mixture model and Burr XII mixture model provide very close estimations.

From Table 6.13, we can see that the estimated values of the shape parameters of the Weibull mixture model are close to the estimates of $\hat{\tau}_j$ of the Burr XII mixture model. This agrees with the relationship (6.17) shown in Watkins'(1999) paper.

Comparing the three best performing mixture models, the Burr XII mixture model has the lowest KS distance; whereas the normal mixture model gives the largest log-likelihood function. However, Burr XII has a total of seven parameters and the estimation procedures are more complicated compared to the other models. The Normal mixture model is not ideal on biological grounds because it allows negative values. On the other hand, the Weibull mixture model has considerably low KS distance and the second largest log-likelihood function. Not to mention that the Weibull mixture model assumes that all observations are strictly positive and the estimation of its parameters are relatively easier compared to the Burr XII mixture model.

In this chapter, not only we have seen how the flexibility given by a mixture model provides significantly better fit to a set of raw data, we also learned that the incubation period can be modelled by a wide variety of mixture models.

Given the relatively small amount of data, we cannot draw strong conclusions about the most appropriate model. It is perhaps surprising that the normal model fits so well, since models that include a left skew are usually preferred for incubation period data. Gravenor *et al.* (2003) also showed that a normal distribution provided a reasonable fit to scrapie incubation periods. Despite this, with a range of other models to choose from, we do not favour the normal model as negative values can result. A larger data set must be used as a more rigorous test of model fit. However, at present we recommend the use of the Weibull model due to the goodness of fit. If sufficient data is available for parameter estimation the Burr XII model is the most flexible and promising.

We also note that due to the cost constraints, prion experiments are often characterised by small sample size, and it is exciting to see that a number of candidate models can be suggested, each of which has been shown to provide interesting insights into the underlying infection process even with limited data. The interesting result is that all passages can be well described by a mixture model of just two distributions. The implication from a purely statistical view point is that there exist two discrete strands of infection. Initially the incubation period is drawn exclusively from type "A", after serial passage there is a mixture (in roughly equal proportions) of type A and "B", and in later passage type B

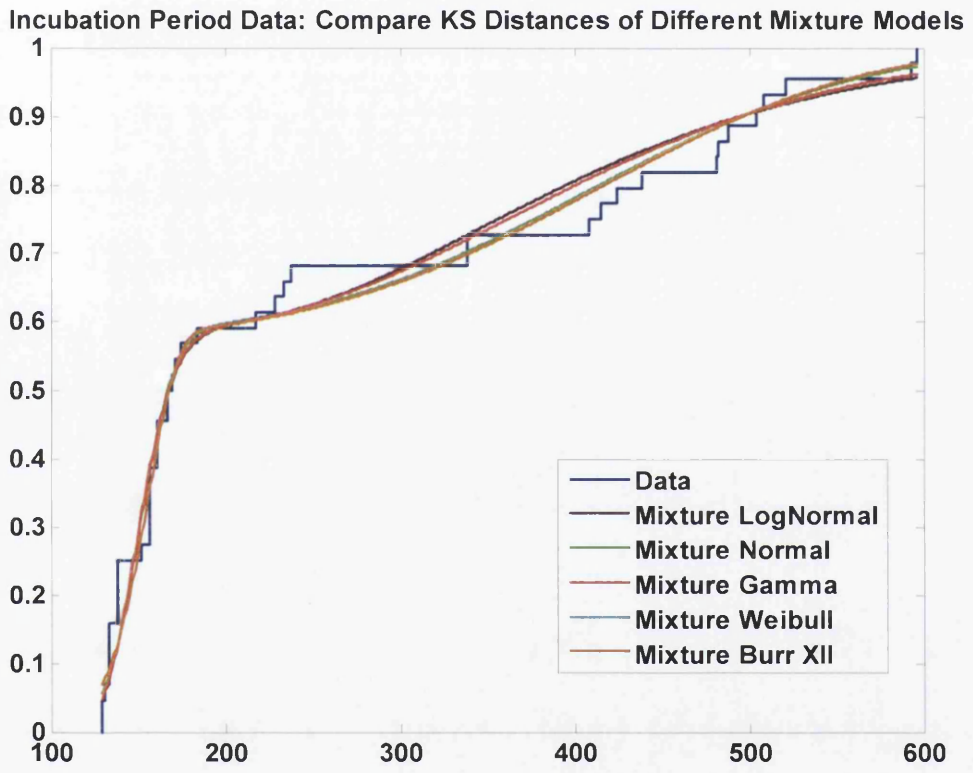


Figure 6.23: Comparison of the ECDF plot of the incubation period data and the fitted CDF plots given by all five mixture models.

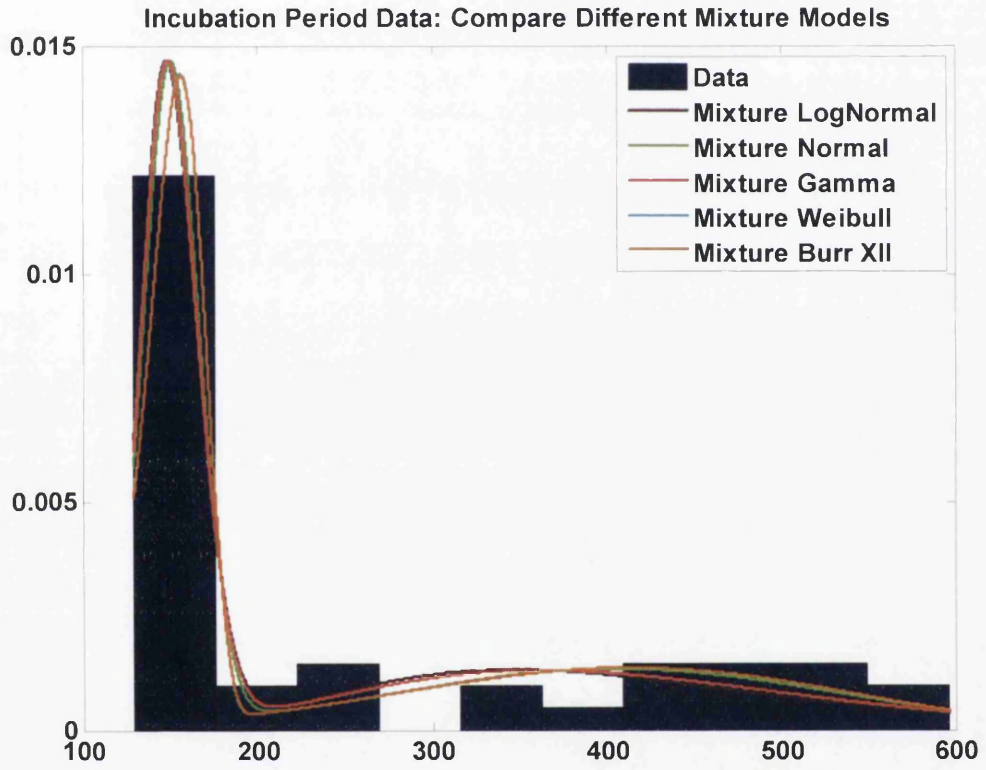


Figure 6.24: Comparison of the EPDF plot of the incubation period data and the fitted PDF plots given by all five mixture models.

Mixture Model	KS	sig	$\boldsymbol{\Theta}$	$l(\boldsymbol{\Theta})$	Summary
Lognormal	0.1223	0.4991	$\mu_1 = 5.0108$ $\sigma_1 = 0.1070$ $\mu_2 = 5.9624$ $\sigma_2 = 0.3430$ $p = 0.5846$	-251.304	$E[T_1] = 150.89$ $Var[T_1] = 262.20$ $E[T_2] = 412.10$ $Var[T_2] = 21205$
Normal	0.1209	0.5145	$\hat{\mu}_1 = 150.74$ $\hat{\sigma}_1 = 15.889$ $\hat{\mu}_2 = 405.29$ $\hat{\sigma}_2 = 126.31$ $\hat{p} = 0.5767$	-249.912	$E[T_1] = 150.74$ $Var[T_1] = 252.46$ $E[T_2] = 405.29$ $Var[T_2] = 15954$
Gamma	0.1231	0.4910	$\hat{\alpha}_1 = 88.998$ $\hat{a} = 0.5903$ $\hat{\alpha}_2 = 9.0820$ $\hat{b} = 0.0223$ $\hat{p} = 0.5805$	-250.627	$E[T_1] = 150.77$ $Var[T_1] = 255.40$ $E[T_2] = 407.63$ $Var[T_2] = 18295$
Weibull	0.1210	0.5132	$\hat{\alpha}_1 = 10.521$ $\hat{a} = 0.0063$ $\hat{\alpha}_2 = 3.7656$ $\hat{b} = 0.0022$ $\hat{p} = 0.5758$	-250.177	$E[T_1] = 150.61$ $Var[T_1] = 298.28$ $E[T_2] = 406.91$ $Var[T_2] = 14534$
Burr XII	0.1205	0.5188	$\hat{\alpha}_1 = 918.48$ $\hat{\tau}_1 = 10.498$ $\hat{a} = 0.0033$ $\hat{\alpha}_2 = 492.21$ $\hat{\tau}_2 = 3.7686$ $\hat{b} = 0.0004$ $\hat{p} = 0.5759$	-250.181	$E[T_1] = 150.37$ $Var[T_1] = 298.79$ $E[T_2] = 405.68$ $Var[T_2] = 14451$

Table 6.13: Comparison of the performances of all mixture models fitted to incubation period data.

comes to dominate and type A has all but disappeared.

We now ask the question of whether this makes sense biologically? We introduced the concept of prion "strains" in Chapter 2. On inspection, the CWD serial passage experiment clearly involved two biologically distinct strains. On autopsy, differences were noticed in the "plaque" size (the build up of aggregates of PrPSc protein) and shape in the brain (Dr Christina Sigurdson, personal communication). One strain was found to develop very dense plaques, while the other developed more diffuse aggregates (the longer incubation period). These patterns were repeatable, and different strains could be "selected" by choosing the donor mouse based on its incubation period. Our statistical results suggest that the mixture models may provide a useful tool for the identification and quantification of prion strains.

We wanted to check if the mixture theory can reveal itself in individual passages and hence we fitted a mixture of two distributions, with known parameters except for the mixing weight p, to each passage. The results are shown from Tables 6.8 to 6.12 alongside the KS distances; clearly our hypothesis does not work. We are thankful that this point has been kindly pointed out by the examiners. Mixture fits very well to the combined data with a respectable data; whereas the sizes of all individual passage datas are very small, for instance, the fifth passage has only a dataset of size five. It is therefore not surprising that the fit is not satisfactory when we fit five data points with a distribution with five parameters. Due to the small number of data available, these negative results might not be very safe. We still believe that if we repeat this when more data points become available for each individual passage in the future, we would find a good fit using a mixture distribution.

Chapter 7

Markov Models for Tracking Sub-Clinical Infection in Serial Passage Prion Studies

According to Prusiner's prion model (1982), diseases such as scrapie, BSE, CWD and CJD are caused by transmissible particles that are devoid of nucleic acid and are composed exclusively of a modified form of a normal host protein, PrP (see Chapter 1). In the previous chapter, we studied the incubation period of prion diseases during serial passage experiments. Here, we use a hidden Markov model framework (introduced in Chapter 1) to study the probability that prion infection is transmitted to an exposed individual in similar experimental systems. Specifically, we raise the possibility that prions may cause sub-clinical infection, that only manifests as disease on subsequent passage when transmitted to a new host, either directly (experimental systems) or via contact (epidemic systems). A sub-clinically infected animal does not show clinical signs of scrapie but can transmit it on to the next generation. Using an experimental system of scrapie disease in mice, the waiting time for a host to exhibit signs of scrapie can be modelled by a special kind of Markov process. Our aim is to "track" the sub-clinically infected animals, giving an estimate of the overall prevalence of clinical scrapie in animals at each generation.

7.1 Aim

The objective of this analysis is to estimate the proportion of animals which are sub-clinically infected in each generation of a serial passage study. We aimed to construct a good model to track the hidden state, "Sub-clinically Infected" , because one sometimes can never tell if a non-diseased mouse is healthy or sub-clinically infected. Different models are constructed to estimate the transition probabilities from the "Sub-clinically Infected" state, and also the initial proportion of the mice which are sub-clinically infected. We introduce a sequence of models of increasing complexity, and study their properties.

7.2 Markov Model

After exposure to recombinant PrP or mouse brain inoculum in subsequent passage, the mice exist in one of three states, $\{D, H, S\}$, which are "Diseased" (definitely infected) (D), "Uninfected and Healthy" (H), and "Sub-clinically Infected" (S). Transition between all states is possible. The states refer to the condition of the mouse at the time it is used to initiate further passage. The matrix of transition probabilities is:

$$\boldsymbol{P} = \begin{bmatrix} 1 - p_{dh} - p_{ds} & p_{dh} & p_{ds} \\ p_{hd} & 1 - p_{hd} - p_{hs} & p_{hs} \\ p_{sd} & p_{sh} & 1 - p_{sd} - p_{sh} \end{bmatrix} \tag{7.1}$$

where

$1 - p_{dh} - p_{ds}$ is the probability of a diseased mouse causing disease on serial transfer (D to D transition)

p_{dh} is the probability of a D to H transition

p_{ds} is the probability of a D to S transition

p_{hd} is the probability of a H to D transition

$1 - p_{hd} - p_{hs}$ is the probability of a H to H transition

p_{hs} is the probability of a H to S transition

p_{sd} is the probability of a S to D transition

p_{sh} is the probability of a S to H transition

$1 - p_{sd} - p_{sh}$ is the probability of a S to S transition.

Note that the transitions from state D to state H and transitions from state D to state S are almost impossible, so we assume that p_{dh} and p_{ds} are close to 0. In practice, the possibility of transitions from state H to state S and state H to state D is also extremely low, so we assume that $1 - p_{hd} - p_{hs}$ is close to 1. Following the assumptions, we let

$$\begin{aligned} p_{dh} &\approx 0, \\ p_{ds} &\approx 0, \\ 1 - p_{dh} - p_{ds} &\approx 1, \\ p_{hd} &\approx 0, \\ p_{hs} &\approx 0, \\ 1 - p_{hd} - p_{hs} &\approx 1. \end{aligned}$$

Hence, the transition matrix in (7.1) becomes

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ p_{sd} & p_{sh} & 1 - p_{sd} - p_{sh} \end{bmatrix}. \tag{7.2}$$

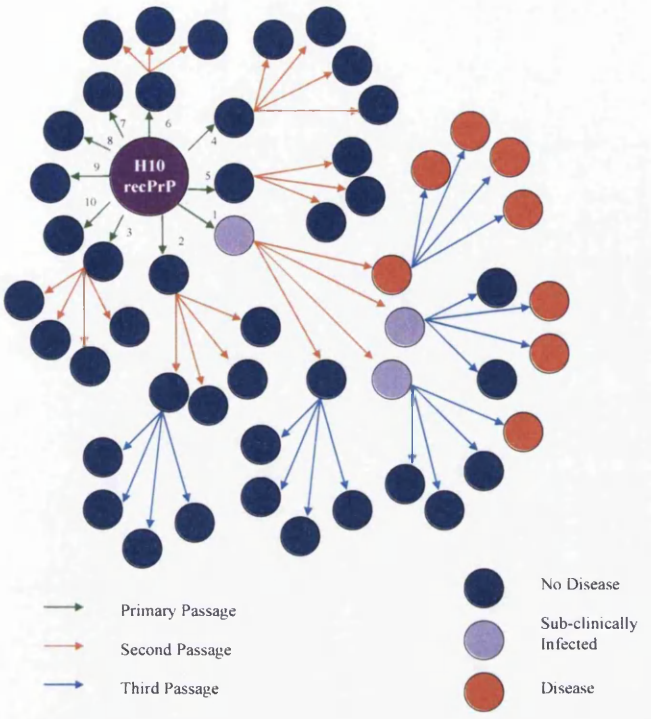


Figure 7.1: H10 recPrP$^{\beta}$ experimental data and design of a typical "serial passage" experiment. Data provided by Professor A. Aguzzi, Institute of Neuropathology, University of Zurich.

7.3 H10 PrP$^{\beta}$ Experimental Data

In the experimental group H10 recPrP$^{\beta}$, 10 mice are, at first passage, exposed to a recombinant PrP sample (see Figure 7.1 for the structure of the serial passages). None of the ten mice develop disease. Six of the mice are chosen at random for second passage into either three or four mice. In one of four mice arising from one of the primary passage animals, scrapie disease is detected, and successfully transmitted to a further four mice (diseased mice are indicated by red circles in Figure 7.1). In the same series, two of the three non-diseased second passage mice (indicated by purple circles) generated diseased animals in third passage. The statistical model is defined as a Markov chain, so we picture the data set as a forest with ten trees bounded with its leaves. The details of the transition between mice are presented in Table 7.1 and Table 7.2 *(since in the experiment we could not tell whether a mouse which shows no clinical symptom is healthy (H) or sub-clinically infected (S), we express the state of such a mouse as "no disease", denoted by E).* In our analysis, we only consider the first six trees because Tree 7 to Tree 10 are not informative, having only one generation. The implication is that recPrP^{Sc} causes a sub-clinical infection on first passage, which can manifest as disease on subsequent passage.

Tree	Root	First	Second
1	S	D	D
			D
			D
			D
		S	E
			D
			D
			E
		S	D
			E
			E
			E
		E	E
			E
			E
			E
2	E	E	E
			E
			E
			E
		E	
		E	
		E	

Table 7.1: H10 recPrP$^{\beta}$ Experimental Group: Tree 1 and Tree 2

Tree	Root	First
3	E	E
		E
		E
		E
4	E	E
		E
		E
		E
5	E	E
		E
		E
6	E	E
		E
		E
7	E	
8	E	
9	E	
10	E	

Table 7.2: H10 recPrP$^{\beta}$ Experimental Group: Tree 3 to Tree 10

7.4 Naïve Two-State Model

An initial analysis of the serial passage PrP^{Sc} experimental data is a two state Markov Chain, where states S and H are grouped as one state, E. Therefore, only two states are considered in this simpler model, which is "Diseased" (D), and "No-Disease " (sub-clinal or uninfected) (E). The transition matrix is

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 \\ p_{ed} & 1-p_{ed} \end{bmatrix}$$

where p_{ed} is the probability of a E to D transition. The starting probabilities are

$$\boldsymbol{\pi}^{(0)} = \begin{bmatrix} \pi_d & \pi_e \end{bmatrix} = \begin{bmatrix} 1-\pi_e & \pi_e \end{bmatrix}.$$

In the experimental data, none of the primary-passage mice showed scrapie disease, so $\boldsymbol{\pi}^{(0)} = \begin{bmatrix} 0 & 1 \end{bmatrix}$. It is informative for us to know the distribution of the state of mice when the number of serial passage increases. Without carrying on the experiments for a large number of passages, the distribution vector, $\boldsymbol{\pi}^{(n)}$ forecasts whether all mice are diseased when the number of passages increases, or there will be a fixed proportion of mice in the future generation which are disease-free. The expected probabilities at the n^{th} passage can be found by

$$\begin{bmatrix} \pi_d & \pi_e \end{bmatrix}^{(n)} = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ p_{ed} & 1-p_{ed} \end{bmatrix}^n.$$

Let $\boldsymbol{P}^n$ denotes the n^{th} transition matrix

$$\boldsymbol{P}^n = \begin{bmatrix} 1 & 0 \\ p_{ed} & 1-p_{ed} \end{bmatrix}^n = \begin{bmatrix} 1 & 0 \\ 1-(1-p_{ed})^n & (1-p_{ed})^n \end{bmatrix}. \tag{7.3}$$

It follows that the expected probabilities at the n^{th} passage are

$$\boldsymbol{\pi}^{(n)} = \begin{bmatrix} \pi_d & \pi_e \end{bmatrix}^{(n)} = \begin{bmatrix} 1-(1-p_{ed})^n & (1-p_{ed})^n \end{bmatrix}.$$

This means that

$$\Pr\left[\text{Diseased at } n^{th} \text{ passage}\right] = 1-(1-p_{ed})^n, \tag{7.4}$$

and

$$\Pr\left[\text{Not diseased at } n^{th} \text{ passage}\right] = (1-p_{ed})^n. \tag{7.5}$$

Note (7.4) is the CDF of a geometric distribution with parameter p_{ed}; while (7.5) is the survival function of a geometric distribution with parameter p_{ed}.

Consider the likelihood

$$L = p_{ed}^{n_d}(1-p_{ed})^{n_e}$$

where n_d is the number of E to D transitions, and n_e is the number of E to E transitions.

Transition	Frequency
$D \rightarrow D$	4
$D \rightarrow E$	0
$E \rightarrow D$	4
$E \rightarrow E$	34
From D	4
From E	38

Table 7.3: Naïve Two-State Model: Number of transitions between states of mice given H10 recPrP$^{\beta}$ on primary passage.

The log-likelihood is

$$l = \log L(p_{ed}) = n_d \log(p_{ed}) + n_e \log(1 - p_{ed}).$$

Equating the score function to zero, we get the ML estimate of p_{ed}:

$$\frac{\partial l}{\partial p_{ed}} = \frac{n_d}{p_{ed}} - \frac{n_e}{1 - p_{ed}} = 0$$

$$\Leftrightarrow p_{ed} = \frac{n_d}{(n_d + n_e)}. \tag{7.6}$$

7.4.1 Results and Discussion

We perform the simplest calculation of p_{ed} in this section by counting the number of transitions between states in one of the experimental data. Table 7.3 shows the number of transitions between the two states.

The likelihood function is given by

$$L = p_{ed}^4 (1 - p_{ed})^{34}.$$

Using (7.6) we have

$$\hat{p}_{ed} = \frac{4}{38}.$$

Hence, the transition matrix for this experimental data is

$$\hat{\boldsymbol{P}} = \begin{bmatrix} 1 & 0 \\ \frac{4}{38} & \frac{34}{38} \end{bmatrix}.$$

Therefore, $p_{ed} = \frac{4}{38}$ is the estimate of the transition probability from state "No-Disease" to state "Diseased". Following (7.3), the transition matrix after n^{th} passage is

$$\hat{\boldsymbol{P}}^n = \begin{bmatrix} 1 & 0 \\ 1 - \left(\frac{34}{38}\right)^n & \left(\frac{34}{38}\right)^n \end{bmatrix}.$$

n	$P[X_n = D \mid X_0 = E]$	$P[X_n = E \mid X_0 = E]$
1	0.1053	0.8947
2	0.1995	0.8006
3	0.2837	0.7163
4	0.3591	0.6409
5	0.4266	0.5734
6	0.4869	0.5131
7	0.5409	0.4591
8	0.5893	0.4107
9	0.6325	0.3675
10	0.6712	0.3288
11	0.7058	0.2942
12	0.7368	0.2632
13	0.7645	0.2355
14	0.7893	0.2107
15	0.8115	0.1886
16	0.8313	0.1687
17	0.8491	0.1510
18	0.8649	0.1351
19	0.8792	0.1208
20	0.8919	0.1081

Table 7.4: Naïve Two-State Model: Transition probabilites of mice in states D and E from state E on passage n.

The transition probabilities from state E after n^{th} passage is given in Table 7.4; while the proportions of mice in state D and E are shown in Table 7.5. We observe that when the number of passages increases, the possibility of a mouse being diseased is higher. According to the two-state model, the probability of mice being diseased approaches to 1 when the number of serial passage increases.

7.5 Naïve Three-State Model

In this section, we consider a three-state Markov Chain, where we consider three states for an exposed mouse: "Disease" (D), "Uninfected and Healthy" (H) and "Sub-clinically Infected" (S). The states refer to the condition of the mouse at the time it is used to initiate further passage. This is a naïve model because we make an assumption that all sub-clinically infected hosts are detectable.

Let π_d, π_h and π_s be the expected proportion of animals in states D, H and S on passage n. The parameters governing the starting conditions are:

$$\boldsymbol{\pi}^{(0)} = \begin{bmatrix} \pi_d & \pi_h & \pi_s \end{bmatrix}.$$

We recall the matrix of transition probabilities for a three-state model (as in (7.2), is a lower

n	π_d	π_e
0	0	1
1	0.1053	0.8947
2	0.1995	0.8006
3	0.2837	0.7163
4	0.3591	0.6409
5	0.4266	0.5734
6	0.4869	0.5131
7	0.5409	0.4591
8	0.5893	0.4107
9	0.6325	0.3675
10	0.6712	0.3288
11	0.7058	0.2942
12	0.7368	0.2632
13	0.7645	0.2355
14	0.7893	0.2107
15	0.8115	0.1886
16	0.8313	0.1687
17	0.8491	0.1510
18	0.8649	0.1351
19	0.8792	0.1208
20	0.8919	0.1081

Table 7.5: Naïve Two-State Model: Proportion of mice in states D and E on passage n.

triangular matrix (where the entries above the main diagonal are zero).

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ p_{sd} & p_{sh} & 1 - p_{sd} - p_{sh} \end{bmatrix},$$

where

p_{sd} is the probability of a S to D transition, and

p_{sh} is the probability of a S to H transition.

The expected probabilities at n^{th} passage can be found by

$$\boldsymbol{\pi}^{(n)} = \begin{bmatrix} \pi_d & \pi_h & \pi_s \end{bmatrix}^{(n)} = \begin{bmatrix} \pi_d & \pi_h & \pi_s \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ p_{sd} & p_{sh} & 1 - p_{sd} - p_{sh} \end{bmatrix}^n,$$

where the n^{th} transition matrix is in the following form:

$$\boldsymbol{P}^n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \dfrac{p_{sd}\left[1 - (1 - p_{sd} - p_{sh})^n\right]}{p_{sd} + p_{sh}} & \dfrac{p_{sh}\left[1 - (1 - p_{sd} - p_{sh})^n\right]}{p_{sd} + p_{sh}} & (1 - p_{sd} - p_{sh})^n \end{bmatrix}. \quad (7.7)$$

Since $\pi_d^{(0)} = 0$

$$\boldsymbol{\pi}^{(n)} = \left[\frac{\pi_s p_{sd}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} \quad \pi_h + \frac{\pi_s p_{sh}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} \quad \pi_s\left(1-p_{sd}-p_{sh}\right)^n \right]. \tag{7.8}$$

The MLE of p_{sd} and p_{sh} are those which maximise the likelihood

$$L = p_{sd}^{n_d} p_{sh}^{n_u} \left(1-p_{sd}-p_{sh}\right)^{n_s},$$

where

n_d is the number of S to D transitions,

n_h is the number of S to H transitions, and

n_s is the number of S to S transitions.

The log-likelihood is

$$l = \ln(L) = n_d \ln(p_{sd}) + n_h \ln(p_{sh}) + n_s \ln\left(1-p_{sd}-p_{sh}\right).$$

Equating the score functions to zero,

$$\begin{aligned} \frac{\partial l}{\partial p_{sd}} &= \frac{n_d}{p_{sd}} - \frac{n_s}{1-p_{sd}-p_{sh}} = 0, \\ \frac{\partial l}{\partial p_{sh}} &= \frac{n_h}{p_{sh}} - \frac{n_s}{1-p_{sd}-p_{sh}} = 0, \end{aligned}$$

we therefore obtain the ML estimate of p_{sd} and p_{sh} :

$$p_{sd} = \frac{n_d\left(1-p_{sh}\right)}{n_d+n_s}, \tag{7.9}$$

$$p_{sh} = \frac{n_h\left(1-p_{sd}\right)}{n_h+n_s}. \tag{7.10}$$

Substitute (7.9) into (7.10), we have

$$p_{sh} = \frac{n_h\left(\dfrac{n_d+n_s-n_d\left(1-p_{sh}\right)}{n_d+n_s}\right)}{n_h+n_s}$$

$$\Leftrightarrow p_{sh} = \frac{n_h}{n_h+n_d+n_s}, \tag{7.11}$$

hence, upon substituting (7.11) into (7.9), we have

$$p_{sd} = \frac{n_d\left(1-\dfrac{n_h n_s}{\left[\left(n_h+n_s\right)\left(n_d+n_s\right)-n_h n_d\right]}\right)}{n_d+n_s}$$

Transition	Frequency
$D \rightarrow D$	4
$D \rightarrow H$	0
$D \rightarrow S$	0
$H \rightarrow D$	0
$H \rightarrow H$	26
$H \rightarrow S$	0
$S \rightarrow D$	4
$S \rightarrow H$	6
$S \rightarrow S$	2
From D	4
From H	26
From S	12

Table 7.6: Naïve Three-State Model: Number of transitions between states of mice given recPrP on primary passage.

$$\Leftrightarrow p_{sd} = \frac{n_d}{n_h + n_d + n_s}. \tag{7.12}$$

In general,

$$p_{ij} = \Pr[X_n = j | X_{n-1} = i] = \frac{\text{Number of transitions from state } i \text{ to state } j}{\text{Number of transitions from state } i}.$$

7.5.1 Results and Discussion

We perform the simplest calculation of p_{sd} and p_{sh} in this section by counting the number of transitions between states in the experimental data. Table 7.6 shows the number of transitions between the three states. The estimates of p_{sh} and p_{sd} can be obtained by following (7.11) and (7.12). As a result,

$$\hat{p}_{sd} = \frac{1}{3},$$

and

$$\hat{p}_{sh} = \frac{1}{2},$$

whereas the transition matrix in (7.2) is estimated to be

$$\hat{\boldsymbol{P}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{bmatrix}.$$

n	$P[X_n = D \mid X_0 = S]$	$P[X_n = H \mid X_0 = S]$	$P[X_n = S \mid X_0 = S]$
1	0.3333	0.5000	0.1667
2	0.3889	0.5833	0.0278
3	0.3982	0.5972	0.0046
4	0.3997	0.5995	0.0008
5	0.4000	0.5999	0.0001
6	0.4	0.6	0
7	0.4	0.6	0
8	0.4	0.6	0
9	0.4	0.6	0
10	0.4	0.6	0
11	0.4	0.6	0
12	0.4	0.6	0
13	0.4	0.6	0
14	0.4	0.6	0
15	0.4	0.6	0
16	0.4	0.6	0
17	0.4	0.6	0
18	0.4	0.6	0
19	0.4	0.6	0
20	0.4	0.6	0

Table 7.7: Naïve Three-State Model: Transition probabilities of mice in states D, S and U from state S on passage n.

Following (7.7), the n^{th} transition matrix is

$$\hat{\boldsymbol{P}}^n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \dfrac{\frac{1}{3}\left[1-\left(\frac{1}{6}\right)^n\right]}{\frac{5}{6}} & \dfrac{\frac{1}{2}\left[1-\left(\frac{1}{6}\right)^n\right]}{\frac{5}{6}} & \left(\frac{1}{6}\right)^n \end{bmatrix}.$$

Table 7.7 shows the predicted transition probabilities from state S (sub-clinically infected) after n^{th} passage.

This naïve model does not estimate the initial proportion of mice which are sub-clinically infected. By making the assumption that sub-clinically infected mice are always observed, we set π_s as $\frac{1}{6}$ because there is one mouse out of six, ignoring trees 7 to 10, in the experimental sample which is sub-clinically infected at the first passage (as shown in Figure 7.1), therefore

$$\hat{\boldsymbol{\pi}}^{(0)} = \begin{bmatrix} \pi_d & \pi_h & \pi_s \end{bmatrix} = \begin{bmatrix} 0 & \frac{5}{6} & \frac{1}{6} \end{bmatrix}.$$

n	π_d	π_h	π_s
0	0	0.8333	0.1667
1	0.0556	0.9167	0.0278
2	0.0648	0.9306	0.0046
3	0.0664	0.9329	0.0008
4	0.0666	0.9333	0.0001
5	0.0667	0.9333	0
6	0.0667	0.9333	0
7	0.0667	0.9333	0
8	0.0667	0.9333	0
9	0.0667	0.9333	0
10	0.0667	0.9333	0
11	0.0667	0.9333	0
12	0.0667	0.9333	0
13	0.0667	0.9333	0
14	0.0667	0.9333	0
15	0.0667	0.9333	0
16	0.0667	0.9333	0
17	0.0667	0.9333	0
18	0.0667	0.9333	0
19	0.0667	0.9333	0
20	0.0667	0.9333	0

Table 7.8: Naïve Two-State Model: Proportion of mice in states D, S and H on passage n.

Then from (7.8), we predict the proportion of mice in each state for all n passage

$$\hat{\boldsymbol{\pi}}^{(n)} = \left[\frac{1}{6}\left(\frac{\frac{1}{3}\left[1-\left(\frac{1}{6}\right)^n\right]}{\frac{5}{6}} \right) \quad \frac{5}{6}+\frac{1}{6}\left(\frac{\frac{1}{2}\left[1-\left(\frac{1}{6}\right)^n\right]}{\frac{5}{6}} \right) \quad \left(\frac{1}{6}\right)^n \right].$$

Table 7.8 shows the distribution of mice in each state on passage n. At 7^{th} passage, the distribution is stationary with the proportion of mice in state D being 0.0667; proportion of mice with no disease as 0.9333; while the proportion of mice being sub-clinically disease being as small as zero.

We know, from Table 7.6, that $n_d = 4, n_h = 6$ and $n_s = 2$. The likelihood function is given by

$$L = p_{sd}^4 p_{sh}^6 \left(1 - p_{sd} - p_{sh}\right)^2.$$

Using (7.9) and (7.10), we have

$$\begin{aligned} \hat{p}_{sd} &= \frac{1}{3}, \\ \hat{p}_{sh} &= \frac{1}{2}. \end{aligned}$$

This is the same result as that obtained from counting the number of transitions.

The main drawback of this naïve model is that all sub-clinically infected mice are assumed to have been identified and the undiseased mice are definitely healthy. However, the reason for constructing this simple model is to have a feeling on how it works when three states are considered. As shown in Table 7.8, after the 5^{th} serial passage, the state distribution becomes stationary where there are 6.7% of the population being diseased; 93.3% of the population being healthy; while no mouse will be sub-clinically infected in the long run.

7.6 Semi-Naïve Three-State Model

In this section, we devise a model that estimates the probabilities in (7.2) with the observations of the time of occurrence of disease. Since this sample is more sophisticated than the two naïve models discussed earlier, and it is not advanced enough to predict the prevalence of scrapie at the initial passage π_s, we name it a semi-naïve model.

From (7.7) we know the transition matrix at the n^{th} passages is

$$\boldsymbol{P}^n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \dfrac{p_{sd}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} & \dfrac{p_{sh}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} & (1-p_{sd}-p_{sh})^n \end{bmatrix}.$$

If we clump the two indistinguishable states, H and S together into a level which is named as E, we have:

$$\begin{aligned} \boldsymbol{P}^n &= \begin{bmatrix} 1 & 0 \\ \begin{bmatrix} 0 \\ \dfrac{p_{sd}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ \dfrac{p_{sh}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} & (1-p_{sd}-p_{sh})^n \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ \boldsymbol{P}_{ED} & \boldsymbol{P}_{EE} \end{bmatrix}, \end{aligned}$$

where $\boldsymbol{P}_{ED}$ is a 2×1 vector and $\boldsymbol{P}_{EE}$ is a 2×2 matrix:

$$\boldsymbol{P}_{ED} = \begin{bmatrix} 0 \\ \dfrac{p_{sd}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} \end{bmatrix},$$

$$\boldsymbol{P}_{EE} = \begin{bmatrix} 1 & 0 \\ \dfrac{p_{sh}\left[1-(1-p_{sd}-p_{sh})^n\right]}{p_{sd}+p_{sh}} & (1-p_{sd}-p_{sh})^n \end{bmatrix},$$

and hence the probability for the successor of a mouse in state E at $(n-1)^{th}$ passage to

have the same state as its predecessor can be calculated by

$$\begin{aligned} \Pr\left[X_n = E | X_{n-1} = E\right] &= \frac{p_{sh} + p_{sd}\left(1 - p_{sd} - p_{sh}\right)^n}{p_{sd} + p_{sh}} \\ &= p + (1-p)\left(1-b\right)^n, \end{aligned} \tag{7.13}$$

whereas the probability for the successor of a mouse in state E at $(n-1)^{th}$ passage to show scrapie disease is given by

$$\begin{aligned} \Pr\left[X_n = D | X_{n-1} = E\right] &= \frac{p_{sd}\left[1 - \left(1 - p_{sd} - p_{sh}\right)^n\right]}{p_{sd} + p_{sh}} \\ &= (1-p)\left[1 - (1-b)^n\right], \end{aligned} \tag{7.14}$$

where we set

$$p = \frac{p_{sh}}{p_{sd} + p_{sh}}, \tag{7.15}$$

and

$$b = p_{sd} + p_{sh}. \tag{7.16}$$

The probability in (7.13) is in the form of the survival function of a mixture of two geometric distributions with the probability of diseased from the first component as 0, and the probability of diseased from the second component as b. As the aim is to study the survival distribution for the mice which are infected by the H10 recombinant prion protein, the mice which are not infected by the recPrP will stay healthy and never experience scrapie disease. Therefore, we let N be the random variable denoting the number of passages of the mice needed to observe a diseased successor. A healthy mouse which is not infected by recPrP will have $N = \infty$, indicating the event that the successors of a healthy mice will never be diseased because of recPrP. This means that the distribution of number of passages to disease is a mixture of two geometric distributions. The survival function of such a distribution is

$$\begin{aligned} S(n) &= \Pr\left[N > n\right] \\ &= p + (1-p)\left(1-b\right)^n \\ &= p + (1-p)\, S_2(b), \end{aligned} \tag{7.17}$$

where

$$S_2(b) = (1-b)^n$$

for $n = 1, 2, ...$ Since a healthy mouse will never show scrapie disease, the probability of diseased for a healthy mouse is zero. This is confirmed by (7.14), where the CDF of N

$$\begin{aligned} F(n) &= \Pr\left[N \leq n\right] \\ &= (1-p)\left[1 - (1-b)^n\right] \end{aligned} \tag{7.18}$$

is only contributed by the second component, which is the cause of interest. Consequently, we propose that N has a mixture of two geometrics with the parameter of the first component being zero and the parameter of the second component as b.

(7.17) is known as Long-Term Survivor Mixture Model, where p represents the fraction of the population which never fail due to the cause of interest. Therefore, $S_1(t) = 1$ for all t. A similar model has been constructed by Ng & McLachlan (1998) to analyse breast cancer data. They obtained a consistent estimator of the conditional survival function for the cause of interest using the modified long-term survival model which is based on a partial ML approach. Their survival function is in the form of

$$p + (1-p) S_2(t; \theta_2)$$

where $S_2(t; \theta)$ is the conditional survival function for death from breast cancer. They made a comparison between the partial and full ML approaches and found good agreement between the two methods.

7.6.1 Estimating Parameters with a Likelihood Involving PMF and a Survival Function

We have found that the probability of a transition from E to E is in the form of survival function of a mixture of two geometric distributions; the probability of a transition from E to D is in the form of cumulative distribution of a mixture of two geometric distributions. Therefore, we assume that the distribution of number of generations until first disease, N, is a mixture of two geometric distributions with PMF

$$f(n) = (1-p) b (1-b)^{n-1}.$$

If we have full data, the parameters of the distribution of N can be estimated by the MLE where the likelihood is

$$L = \prod_{i=i}^{n_o} f(n_i).$$

However, the experimental data contains censored data. The contribution of each diseased subject to the likelihood is the PMF for the event of disease at the passage that scrapie disease is detected in a successor. The contribution for each mice which show no disease at n^{th} generation is simply the probability of surviving, which is denoted by $S(n)$. Therefore, for incomplete data, the likelihood function contains both the survival function and the probability distribution function:

$$\begin{aligned} L &= \prod_{i \in E} \left(p + (1-p)(1-b)^{n_i}\right) \prod_{i \in D} (1-p) b (1-b)^{n_i - 1} \qquad (7.19) \\ &= (1-p)^d b^d \prod_{i \in E} \left(p + (1-p)(1-b)^{n_i}\right) \prod_{i \in D} (1-b)^{n_i - 1}, \end{aligned}$$

Waiting Time	Observed Data
n_d	2 3 3 3
n_e	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3

Table 7.9: Semi Naïve Three-State Model: Number of passage until event.

where d is the number of diseased mouse. The log-likelihood is therefore

$$\begin{aligned} l &= \ln L \qquad (7.20) \\ &= d\ln(1-p) + d\ln b + \sum_{i\in E} \ln\left(p + (1-p)(1-b)^{n_i}\right) + \sum_{i\in D} (n_i - 1)\ln(1-b). \end{aligned}$$

The MLE of b and p can be obtained by maximising (7.20); whereas the estimates of p_{sh} and p_{sd} can be found as follows:

$$p_{sh} = pb \qquad (7.21)$$

and

$$p_{sd} = (1-p)\, b. \qquad (7.22)$$

Therefore, the transition probabilities, p_{sd} and p_{sh} are found by maximising the log-likelihood function in (7.20).

7.6.2 Results and Discussion

In this section, we demonstrate the application of the semi-naïve model on the experimental data, where the discrete waiting time in states D and E are given in Table 7.9.

Therefore (7.20) is in the following form:

$$l = 4\ln(1-p) + 4\ln b + 17\ln\left(p + (1-p)(1-b)^2\right) + 13\ln\left(p + (1-p)(1-b)^3\right) + 7\ln(1-b).$$

Using Mathematica to maximise the likelihood numerically, the MLE's of p and b are

$$\begin{aligned} \hat{p} &= 4.09 \times 10^{-7}, \\ \hat{b} &= 0.04762, \end{aligned}$$

and hence p_{sd} and p_{sh} are estimated as

$$\begin{aligned} \hat{p}_{sd} &= 0.04762, \\ \hat{p}_{sh} &= 1.95 \times 10^{-8}, \\ \hat{p}_{ss} &= 1 - \hat{p}_{sd} - \hat{p}_{sh} = 0.95238. \end{aligned}$$

n	$P[X_n = D \mid X_0 = S]$	$P[X_n = H \mid X_0 = S]$	$P[X_n = S \mid X_0 = S]$
1	0.0476	1.95×10^{-8}	0.9524
2	0.0930	3.80×10^{-8}	0.9070
3	0.1362	5.57×10^{-8}	0.8638
4	0.1773	7.25×10^{-8}	0.8227
5	0.2165	8.85×10^{-8}	0.7835
6	0.2538	1.04×10^{-7}	0.7462
7	0.2893	1.18×10^{-7}	0.7107
8	0.3232	1.32×10^{-7}	0.6768
9	0.3554	1.45×10^{-7}	0.6446
10	0.3861	1.58×10^{-7}	0.6139
11	0.4153	1.70×10^{-7}	0.5847
12	0.4432	1.81×10^{-7}	0.5568
13	0.4697	1.92×10^{-7}	0.5303
14	0.4949	2.02×10^{-7}	0.5051
15	0.5190	2.12×10^{-7}	0.4810
16	0.5419	2.22×10^{-7}	0.4581
17	0.5637	2.31×10^{-7}	0.4363
18	0.5845	2.39×10^{-7}	0.4155
19	0.6043	2.47×10^{-7}	0.3957
20	0.6231	2.55×10^{-7}	0.3769

Table 7.10: Semi Naïve Three-State Model: Transition probabilities of mice in states D, S and H from state S on passage n.

In words, there is 4% chance for the successor of a mouse which is sub-clinically infected at the primary passage to show scrapie disease. The chance for the successor of a sub-clinically infected mouse to be healthy is as slim as 0%, i.e. it is almost certain that sub-clinically infected mice will pass on infection, in which 95% of the "contacts" will remain sub-clinical.

Table 7.10 shows the transition probabilities from state S (sub-clinically infected) at n^{th} passage. Since this model does not provide us an estimate of the initial proportion of state $\boldsymbol{\pi}^{(0)}$, we shall make a crude estimate of $\hat{\pi}_s = \frac{1}{6}$ and present the estimates of $\boldsymbol{\pi}^{(n)}$ given by (7.8) in Table 7.11. As seen in Table 7.10, given that a mice at initial passage is sub-clinically infected, the probability for its successor to show scrapie disease increases with the number of passages; conversely, it will become less likely for the later passage mice to be sub-clinically infected. According to this model, the chance for the successor of a sub-clinically infected mouse to stay healthy is nearly impossible. Studying Table 7.11, we found that in the long run the proportion of healthy mice is fixed at 83.3% for any passage. However, there is an increasing proportion for diseased mice; at the same time, less proportion of mice are sub-clinically infected by the prion and $\pi_s^{(n)}$ converges to zero when $n \to \infty$.

n	π_d	π_h	π_s
1	0	0.8333	0.1667
2	0.0079	0.8333	0.1587
3	0.0155	0.8333	0.1512
4	0.0227	0.8333	0.1440
5	0.0296	0.8333	0.1371
6	0.0361	0.8333	0.1306
7	0.0423	0.8333	0.1244
8	0.0482	0.8333	0.1185
9	0.0539	0.8333	0.1128
10	0.0592	0.8333	0.1074
11	0.0644	0.8333	0.1023
12	0.0692	0.8333	0.0975
13	0.0739	0.8333	0.0928
14	0.0783	0.8333	0.0884
15	0.0825	0.8333	0.0842
16	0.0865	0.8333	0.0802
17	0.0903	0.8333	0.0764
18	0.0940	0.8333	0.0727
19	0.0974	0.8333	0.0693
20	0.1007	0.8333	0.0660

Table 7.11: Semi Naïve Three-State Model: Proportion of mice in states D, S and H on passage n.

7.7 Self-Revealing Aggregated Markov Processes on Trees

7.7.1 Introduction

Jalali (2008c) constructed a model for the analysis of serial passage prion data and named it Self-Revealing Aggregated Markov Processes on Trees (SRAMPT). He investigated, in discrete as well as continuous time, homogenous Markov processes with finite state space unfolding a finitely branching tree. Initially, at a node σ, the state of the Markov process may not be completely known. We only know that the state is in a subset $S_\sigma \equiv S_\sigma(\sigma) \neq \varnothing$ of all states. But, as the process branches out from this node, the information revealed on any successor node τ to σ may provide further information about σ. This formally means that at node τ, we generally know that the state at node σ was in the $S_\sigma(\tau) \neq \varnothing$, where $S_\sigma(\tau) \subseteq S_\sigma$. This refinement can continue until this subset reduces to a singleton, in which case the state of σ will be completely known.

7.7.2 A Simple Model

Suppose we have a three-state SRAMPT X in discrete time. We denote the states by D, H and S. State D, which is observable, is a sink state, and if any node is in that state, all the successor nodes will also be in that state, therefore we do not study these successors any

longer. In other words, we truncate our tree at D nodes. State H and S, are not directly observable: at node σ we can only know that the state is D or it is not D. In other words, at any node σ,

$$\begin{aligned} S_\sigma &= \{D\}, \\ \text{or } S_\sigma &= \{H, S\}. \end{aligned}$$

However, if any successor τ of σ is found in state D, and if $S_\sigma = \{H, S\}$, then we know that $S_\sigma(\tau) = \{S\}$. State H never reveals itself in finite time. We assume that transition between states, from one generation to next generation, are Markovian. We make this statement precise by first introducing some notations.

We denote the nodes by a sequence of letters from our alphabets: $\{x, y, ...\}$, and we denote sequences by $\sigma, \tau, ...$etc. These sequences can also be written explicitly, for instance σ may denote the sequence xy. When we write these in explicit form, we just concatenate letters from left to right. At the bottom of our tree we have the empty sequence denoted by $\langle\rangle$.

A node may not have a successor. In this case we call such a node a leaf. All nodes in state D are leaves. Nodes not in state D may have one or more immediate successors. If σ is the sequence corresponding to a node, an immediate successor of this node should be denoted by a sequence of the form σx. We often deliberately confuse a node with the sequence denoting it. If a node τ is a successor of node σ we denote this fact by $\tau \succ \sigma$. If τ is an immediate successor to σ, we write $\tau . \succ \sigma$. We denote by $suc(\sigma)$, the set of all immediate successors of σ. If σ is a leaf, this set is an empty set.

Now we are ready to introduce transition probabilities

$$p_{ij} = \Pr[X(\sigma x) = j | X(\sigma) = i],$$

where $X(\sigma)$ is the state of the process at node σ. Note that this actual set may be hidden to us. We furthermore assume that

1. *Given $X(\sigma)$, $X(\sigma x)$ and $X(\sigma y)$, $x \neq y$, are independent random variables.*
2. *Given $X(\sigma)$ and $X(\tau)$, $\sigma \neq \tau$, $X(\sigma x)$ and $X(\tau y)$ are independent random variables.*

In other words various transitions are completely independent from each other.

It is assumed that in our simple model the matrix of transition probabilities is in the form of (7.2):

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ p_{sd} & p_{sh} & 1 - p_{sd} - p_{sh} \end{bmatrix}.$$

7.7.3 The Likelihood Function of the Simple Model

Suppose we have observed tree T. In finite time our tree can only be finite, so the tree is bounded by its leaves. Some of these leaves may be in state D and some not so. If a leaf is not in state D, then we do not know its exact state. In order to write the likelihood function, we first let the information and D leaves percolate down the tree. So the state of all the predecessors of any D will be known as state S. After percolation of information the set of possible states of σ is denoted by $\bar{S}_\sigma$. In our simple model this set is either $\{D\}$ (if σ is in D state), $\{S\}$ (if σ is not in D state but has a successor in this state), or $\{H, S\}$ (if neither σ nor any of its successors are in state D).

Next we start writing the likelihood function from the leaves to the root of the tree. We do this by induction.

We denote $L_\sigma^{(i)}$ the likelihood beyond σ when this node is in state i. We observe that $L_\sigma^{(i)} = 1$ if σ is a leaf. Now suppose the likelihood function beyond all the immediate successors of σ are obtained and we want to find the likelihood function beyond σ. Then

$$L_\sigma^{(i)} = \prod_{\tau \in suc(\sigma)} \sum_{j \in \bar{S}_\tau} p_{ij} L_\tau^{(j)}. \tag{7.23}$$

Finally, suppose we have found the likelihood function beyond all the root of the tree and we want to find the overall likelihood function. Here we assume that the vector of our initial probabilities is

$$\boldsymbol{\pi}^{(0)} = \begin{bmatrix} 0 & 1-\pi_s & \pi_s \end{bmatrix}.$$

This in particular means that $X(\langle\rangle) \neq 0$. We then have:

$$\begin{aligned} L_{\langle\rangle} &= \pi_s L_{\langle\rangle}^{(S)} && \text{if } \bar{S}_{<>} = \{S\} \\ L_{\langle\rangle} &= (1-\pi_s) L_{\langle\rangle}^{(H)} + \pi_s L_{\langle\rangle}^{(S)} && \text{if } \bar{S}_{<>} = \{H, S\}. \end{aligned} \tag{7.24}$$

If we have a number of independent experiments as above, then the situation can be depicted by the same number of disconnected trees, or a forest. To find the likelihood function of a forest we simply multiply the likelihood functions of its constituent trees.

7.7.4 Results and Discussion

In the experimental group H10 recPrP$^\beta$, ten mice are given at first passage a recombinant PrP or mouse brain inoculum. We now use the SRAMPT model to estimate the relevant parameters, including the initial proportion of mice which are sub-clinically infected, denoted by π. We picture the data as a forest with six trees, as shown in Tables 7.1 and 7.2. Following

(7.23) and (7.24) , the likelihood of each tree is given by

$$\begin{aligned} L_1 &= \pi_s p_{sd}\left[(1-p_{sd}-p_{sh})\,p_{sd}^2\,(1-p_{sd})^2\right] \\ &\quad \times\left[(1-p_{sd}-p_{sh})\,p_{sd}\,(1-p_{sd})^3\right]\left[p_{sh}+(1-p_{sd}-p_{sh})\,(1-p_{sd})^4\right] \\ &= \pi_s p_{sd}^4\,(1-p_{sd}-p_{sh})^2\,(1-p_{sd})^5\left[p_{sh}+(1-p_{sd}-p_{sh})\,(1-p_{sd})^4\right], \end{aligned}$$

$$L_2 = 1-\pi_s+\pi_s\,(1-p_{sd})^3\left[p_{sh}+(1-p_{sd}-p_{sh})\,(1-p_{sd})^4\right],$$

$$L_3 = L_4 = 1-\pi_s+\pi_s\,(1-p_{sd})^4,$$

and

$$L_5 = L_6 = 1-\pi_s+\pi_s\,(1-p_{sd})^3.$$

The ultimate likelihood function of the forest is the product of its constituent trees' likelihood functions:

$$\begin{aligned} L &= \prod_{i=1}^{6} L_i \\ &= \begin{pmatrix} \pi_s p_{sd}^4\,(1-p_{sd}-p_{sh})^2\,(1-p_{sd})^5\left[p_{sh}+(1-p_{sd}-p_{sh})\,(1-p_{sd})^4\right] \\ \left[1-\pi_s+\pi_s\,(1-p_{sd})^3\left[p_{sh}+(1-p_{sd}-p_{sh})\,(1-p_{sd})^4\right]\right] \\ \left[1-\pi_s+\pi_s\,(1-p_{sd})^4\right]^2\left[1-\pi_s+\pi_s\,(1-p_{sd})^3\right]^2 \end{pmatrix}. \end{aligned}$$

Using Mathematica, we find the MLE of p_{sd}, p_{sh} and π_s numerically. The results are as follows:

$$\begin{aligned} \hat{p}_{sd} &= 0.22011, \\ \hat{p}_{sh} &= 4\times 10^{-5}, \\ \hat{p}_{ss} &= 1-\hat{p}_{sd}-\hat{p}_{sh} = 0.77985, \\ \hat{\pi}_s &= 0.26032. \end{aligned}$$

In words, there is a 22% chance for the "contact" of a sub-clinically infected mouse to develop disease. Sub-clinically infected hosts are almost certain to pass on infection, although most will remain sub-clinical. The prevalence of scrapie at initial passage is 26%, far higher than the naïve estimate (17%).

With these SRAMPT estimates, the n^{th} transition probabilities given the initial passage mouse is sub-clinically infected are shown in Table 7.12. When the number of serial passage increases, the probability of a successor, whose predecessor is sub-clinically infected at initial passage, to be infected by scrapie disease increases. Conversely, the chance for the successor of a sub-clinically infected mouse (at initial passage) to stay in state S decreases with the number of passages. Although the SRAMPT estimate of p_{sh} is as low as 0.00004, this

n	$P[X_n = D\|X_0 = S]$	$P[X_n = H\|X_0 = S]$	$P[X_n = S\|X_0 = S]$
1	0.22011	0.00004	0.77985
2	0.39176	0.00007	0.60817
3	0.52563	0.00010	0.47428
4	0.63002	0.00011	0.36986
5	0.71143	0.00013	0.28844
6	0.77492	0.00014	0.22494
7	0.82443	0.00015	0.17542
8	0.86304	0.00016	0.13680
9	0.89315	0.00016	0.10668
10	0.91664	0.00017	0.08320
11	0.93495	0.00017	0.06488
12	0.94923	0.00017	0.05060
13	0.96037	0.00017	0.03946
14	0.96905	0.00018	0.03077
15	0.97583	0.00018	0.02400
16	0.98111	0.00018	0.01871
17	0.98523	0.00018	0.01459
18	0.98844	0.00018	0.01138
19	0.99094	0.00018	0.00888
20	0.99290	0.00018	0.00692

Table 7.12: SRAMPT Model: Transition probabilities of mice in states D, S and H from state S on passage n.

probability does increase at a very slow rate when the number of passages increases. From Table 7.13, we present the SRAMPT forecast of the distribution of state at n^{th} passage. According to the model, it is predicted that the hidden number of sub-clinically infected mice decreases when the number of serial passage increases; and more mice show scrapie disease. Eventually, none of the mice are sub-clinically infected. The proportion of healthy mice is around 74% for all n.

7.8 Discussion and Summary

In this chapter we focused on an important element of prion biology, that of sub-clinical infection. Prion diseases are known to have very long incubation periods. Once infected, an animal may be able to pass on the disease well before any symptoms have alerted us to the infectious state. This factor lead to the early exponential spread of the BSE agent and the 'mad cow disease' epidemic in the UK, when infected but asymptomatic infected cattle were included in cattle feed. The infected status can in many cases be detected experimentally, however this usually relies on post mortem investigations (biochemical detection of PrP protein in the brain). There is a further complication of sub-clinical infection. In some cases, infection cannot even be detected experimentally. The sub-clinical status is then very prolonged, and disease may never be observed in a natural lifespan, even though tissue

n	π_d	π_h	π_s
0	0	0.7397	0.2603
1	0.0573	0.7397	0.2030
2	0.1020	0.7397	0.1583
3	0.1368	0.7397	0.1235
4	0.1640	0.7397	0.0963
5	0.1852	0.7397	0.0751
6	0.2017	0.7397	0.0586
7	0.2146	0.7397	0.0457
8	0.2247	0.7397	0.0356
9	0.2325	0.7397	0.0278
10	0.2386	0.7397	0.0217
11	0.2434	0.7397	0.0169
12	0.2471	0.7397	0.0132
13	0.2500	0.7397	0.0103
14	0.2523	0.7397	0.0080
15	0.2540	0.7397	0.0063
16	0.2554	0.7397	0.0049
17	0.2565	0.7397	0.0038
18	0.2573	0.7397	0.0030
19	0.2580	0.7397	0.0023
20	0.2585	0.7397	0.0018

Table 7.13: SRAMPT Model: Proportion of mice in states D, S and H on passage n.

Model	p_{sd}	p_{sh}	$\hat{\pi}_s$	$\hat{\pi}_d^{(\infty)}$	$\hat{\pi}_h^{(\infty)}$	$\hat{\pi}_s^{(\infty)}$
Naïve 2-state	$p_{ed} = 0.1053$	-	$\hat{\pi}_e = 1$	1	$\hat{\pi}_e^{(\infty)} = 0$	
Naïve 3-state	0.3333	0.5	0.1667	0.0667	0.9333	0
Semi-Naïve	0.0476	1.95×10^{-8}	0.1667	0.1667	0.8333	0
SRAMPT	0.2201	4.00×10^{-5}	0.2603	0.2603	0.7397	0

Table 7.14: Comparison of all Markov models for tracking sub-clinical infection in serial passage prion studies

from a sub-clinical case may be transmitted if any animals are exposed to it (Race *et al.* (2001) and Race *et al.* (2002)). In prion studies there is usually a distinction between a 'pre-clinical' infection (simply an infected animal that has not reached its incubation period) and sub-clinical, which is even more prolonged and may be undetectable even experimentally. However these differences are largely semantic. Sub-clinical infection for diseases such as vCJD raise important questions for screening studies (Hill & Collinge (2003)), most notably for the possibility of iatrogenic contamination (blood transfusions, surgical instruments and other surgical procedures).

We summarise the estimation results provided by each model considered in Table 7.14. It is clear that no mouse will be sub-clinically infected in the long run. Both of the naïve models are obviously not preferred due to their assumption which states that all sub-clinical

infected mice are detectable. The semi-naïve model is not favoured because it does not tell us the initial proportion of sub-clinically infected mice. However, there are close agreements between the semi-naïve model and SRAMPT. Both models predict that it is almost impossible for the successor of a sub-clinically infected mouse to be completely clear of infection. SRAMPT is the most outstanding model because it is the only model that provides a promising estimate of the proportion of sub-clinical infection which may be undetectable.

The differing results drawn from the sequences of models all having the same 'aim' is interesting, and highlights the importance of defining an appropriate model for a biological system, as well as thoroughly exploring its behaviour. In two-state model all subjects eventually exhibit disease, in the naïve and semi-naïve three-state models an equilibrium of the proportion diseased and healthy is reached, but with very different proportions and time scales. The semi-naïve model, based on the mixture distribution for the waiting time is promising, but we favour the SRAMPT model for fully tracking the expected proportion of sub-clinical infections. This model predicts an equilibrium proportion of diseased and healthy mice of 26% and 74% respectively, one that is reached after approximately 14 generations. Most importantly it predicts an initial proportion of 26% sub-clinical infection, far higher than that expected from simple inspection of the data.

We believe the SRAMPT model to be very promising tool that could be applied to a wide range of biological systems. A further extension to the model is explored in the next chapter.

Chapter 8

Application of the Serial Passage Model to the Problem of Sub-clinical Infection in Epidemiological Chains

In this chapter, we consider an interesting further application of the SRAMPT model, which was initially developed for studying the serial passage studies of scrapie. Again, we have three states (D, H, S) which are "Diseased" (infected and showing symptoms), "Healthy" (Uninfected), and "Sub-clinically Infected" (no symptoms are displayed, but infected and potentially infectious). We now consider the Markov model to represent a chain of contacts of an infectious disease, where each individual (node) has a contact with a preceding individual (node) who may or may not be infectious.

The model is conceptually equivalent to the serial passage study, but aims to represent the situation of transmission of a (directly transmitted) infectious disease such as a virus or bacteria (such as measles, influenza or tuberculosis). Instead of the experimental transfer of infection (as in Chapter 7) contacts in the epidemiology model will be determined by behaviour of a population. Contact chains are often constructed by public health epidemiologists during an outbreak to help identify the source of an outbreak and to identify populations at risk. As before, we are primarily interested in the case of sub-clinical infection, which can be highlighted in epidemiology when disease is found in an individual who does not have direct contact with a clinical case, but was instead infected by an intermediate contact who remained asymptomatic. Unlike the scrapie model, we must relax some of the assumptions of the transition matrix. Specifically, we must acknowledge that D is not an absorbing state and transitions from D to H or S are possible (and indeed might be the norm). That is, in the epidemiology case, contact with a diseased individual does not always result in transfer of infection. H of course remains an absorbing state. The transition matrix for the

epidemiology model is

$$\boldsymbol{P} = \begin{bmatrix} p_{dd} & p_{dh} & p_{ds} \\ 0 & 1 & 0 \\ p_{sd} & p_{sh} & p_{ss} \end{bmatrix}. \tag{8.1}$$

Since $p_{dh} = 1 - p_{dd} - p_{ds}$ and $p_{sh} = 1 - p_{sd} - p_{ss}$, the model only has four independent parameters.

In this chapter we show that the small modification to the model leads to considerable complexities in parameter estimation. Using the SRAMPT framework Jalali (2008c) has provided solutions to the parameter estimation problem. Here we apply his results and explore their performance under a range of epidemiological scenarios.

8.1 Maximum Likelihood Estimation

In this section, we explain how SRAMPT can be used to solve the estimation problem of the transition probabilities in the epidemiology model. In his unpublished paper, Jalali (2008c) looked at this problem and showed how the MLE of the probabilities can be obtained. Before we show our simulation studies in a later section, let us first understand how Jalali's solution should be carried out for such a model. The following theoretical exposition follows his paper.

Like the scrapie model, we denote by E, the disjunction of states H and S, when we do not know in which of them our chain is. As it is now possible to go from state D to state H and S, reaching D does not stop us from going further: beyond this state there are many possibilities. Suppose we have a chain of length $n_o + 1$. Unless the chain consists entirely of E's, after the filtering down of information in state D, all states preceding any state D are either D or S. There is, therefore a last state D, after which we have D, H or more E's. Suppose at the root of our chain is a D state, then we have a sequence of D's and S's, ending with a D, called the tail of the chain, followed by k E's, called the head of the chain. Note that k can be any non-negative integer, including zero. For illustration, see Figure 8.1 where $k = 5$.

We use the SRAMPT to estimate these parameters and the likelihood function can be computed by the same inductive method as before. For simplicity, we assume that there is no multiple branching at nodes, thus we study the problem by looking at a sequence or a chain of states.

If $k = 0$, then our chain entirely consists of D's and S's. The likelihood, in this case is

$$L = p_{dd}^{n_{dd}} p_{ds}^{n_{ds}} p_{sd}^{n_{sd}} p_{ss}^{n_{ss}} \tag{8.2}$$

where n_{ij} denotes the number of transitions from state i to j. As increasing probabilities p_{dh} and p_{sh} decreases the likelihood, for obtaining maximum likelihood, we need to set these probabilities equal to zero. This means that $p_{ds} = 1 - p_{dd}$ and $p_{sd} = 1 - p_{ss}$. Hence, in this

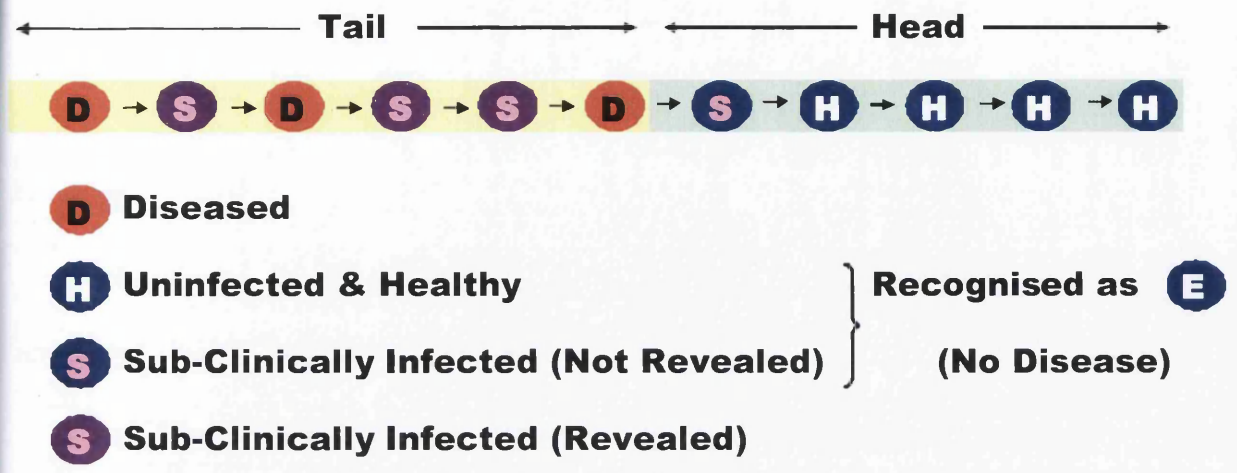


Figure 8.1: The Epidemiology Model: An illustration of a sequence of states.

case the MLE of the parameters are

$$\begin{aligned} \hat{p}_{dd} &= \frac{n_{dd}}{n_{dd}+n_{ds}}, \\ \hat{p}_{ds} &= \frac{n_{ds}}{n_{dd}+n_{ds}}, \\ \hat{p}_{sd} &= \frac{n_{sd}}{n_{ss}+n_{sd}}, \\ \hat{p}_{ss} &= \frac{n_{ss}}{n_{ss}+n_{sd}}. \end{aligned} \tag{8.3}$$

Note that when the chain starts with a D state, $n_{ds} = n_{sd}$.

Now if $k > 0$, then the likelihood function is in the form of

$$L = p_{dd}^{n_{dd}} p_{ds}^{n_{ds}} p_{sd}^{n_{sd}} p_{ss}^{n_{ss}} \left[\begin{bmatrix} p_{dh} & p_{ds} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ p_{sh} & p_{ss} \end{bmatrix}^{k-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right]. \tag{8.4}$$

The matrix $\begin{bmatrix} 1 & 0 \\ p_{sh} & p_{ss} \end{bmatrix}^{k-1}$ has the simple expression $\begin{bmatrix} 1 & 0 \\ y_{k-1} & p_{ss}^{k-1} \end{bmatrix}$. Clearly,

$$\begin{aligned} \begin{bmatrix} 1 & 0 \\ y_k & p_{ss}^k \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ y_{k-1} & p_{ss}^{k-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ p_{sh} & p_{ss} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ y_{k-1} + p_{sh} p_{ss}^{k-1} & p_{ss}^k \end{bmatrix}. \end{aligned}$$

Hence we have the difference equation

$$y_k - y_{k-1} = p_{sh} p_{ss}^{k-1}$$

with the initial condition $y_0 = 0$, which has the solution

$$y_k = p_{sh}\frac{1-p_{ss}^k}{1-p_{ss}}.$$

Hence

$$\begin{bmatrix} 1 & 0 \\ p_{sh} & p_{ss} \end{bmatrix}^k = \begin{bmatrix} 1 & 0 \\ p_{sh}\dfrac{1-p_{ss}^k}{1-p_{ss}} & p_{ss}^k \end{bmatrix}, \tag{8.5}$$

and the likelihood (8.4), after some routine simplification, is

$$L = p_{dd}^{n_{dd}} p_{ds}^{n_{ds}} p_{sd}^{n_{sd}} p_{ss}^{n_{ss}} \left(1 - p_{dd} - p_{ds}p_{sd}\frac{1-p_{ss}^{k-1}}{1-p_{ss}}\right). \tag{8.6}$$

As $n_{ds} = n_{sd}$, we set

$$x = p_{ds}p_{sd}. \tag{8.7}$$

This reduces by one the number of our parameters. It also shows that as far as the product $p_{ds}p_{sd}$ remains the same, the likelihood does not change if p_{ds} or p_{sd} changes. With the new parameter, (8.6) is

$$L = p_{dd}^{n_{dd}} x^{n_{ds}} p_{ss}^{n_{ss}} \left(1 - p_{dd} - x\frac{1-p_{ss}^{k-1}}{1-p_{ss}}\right). \tag{8.8}$$

Setting the derivatives of the log-likelihood with respect to p_{dd} and x equal to zero gives us the following two equations

$$\frac{n_{dd}}{p_{dd}} = \frac{1}{1-p_{dd}-xF(p_{ss})} \Rightarrow n_{dd} = \frac{p_{dd}}{1-p_{dd}-xF(p_{ss})}, \tag{8.9}$$

$$\frac{n_{ds}}{x} = \frac{F(p_{ss})}{1-p_{dd}-xF(p_{ss})} \Rightarrow n_{ds} = \frac{xF(p_{ss})}{1-p_{dd}-xF(p_{ss})}, \tag{8.10}$$

where we let

$$F(p_{ss}) = \frac{1-p_{ss}^{k-1}}{1-p_{ss}}. \tag{8.11}$$

Upon adding (8.9) and (8.10) we have

$$n_{dd} + n_{ds} + 1 = \frac{1}{1-p_{dd}-xF(p_{ss})}, \tag{8.12}$$

and by substituting (8.12) into (8.9), we get

$$\hat{p}_{dd} = \frac{n_{dd}}{n_{dd}+n_{ds}+1}, \tag{8.13}$$

and by putting (8.13) into (8.12), we have

$$xF(p_{ss}) = \frac{n_{ds}}{n_{dd}+n_{ds}+1}. \tag{8.14}$$

Upon setting the derivative of log-likelihood with respect to p_{ss} to zero,

$$\frac{n_{ss}}{p_{ss}} = \frac{xF'(p_{ss})}{1 - p_{dd} - xF(p_{ss})}, \tag{8.15}$$

where F' is the derivative of F. Eliminating x between (8.14) and (8.15) leads to

$$\frac{F'(p_{ss})}{F(p_{ss})} = \frac{n_{ss}}{n_{ds}p_{ss}}. \tag{8.16}$$

But we know that

$$\frac{F'(p_{ss})}{F(p_{ss})} = \frac{d}{dp_{ss}} \ln F(p_{ss}) = \frac{1}{1 - p_{ss}} - \frac{(k-1)\,p_{ss}^{k-2}}{1 - p_{ss}^{k-1}}. \tag{8.17}$$

Hence, the MLE equation for p_{ss} reduces to what Jalali (2008c) called the *Fundamental Equation*

$$\frac{p_{ss}}{1 - p_{ss}} - \frac{(k-1)\,p_{ss}^{k-1}}{1 - p_{ss}^{k-1}} = \frac{n_{ss}}{n_{ds}}. \tag{8.18}$$

8.1.1 Some Special Cases

We have already dealt with the case $k = 0$, now let us study the MLE of the parameters for different lengths of the head of the chain.

1. $k = 1$:

   In this case $F(p_{ss}) = 0$, and the likelihood function (8.8) becomes

   $$L = p_{dd}^{n_{dd}} x^{n_{ds}} p_{ss}^{n_{ss}} (1 - p_{dd}). \tag{8.19}$$

   As this does not depend on p_{dh} and p_{sh}, and they both contribute negatively to p_{dd}, p_{ds} and x, then

   $$\hat{p}_{dh} = \hat{p}_{sh} = 0,$$

   and the likelihood (8.19) is reduced to

   $$L = p_{dd}^{n_{dd}} (1 - p_{dd})^{n_{ds}+1} (1 - p_{ss})^{n_{ds}} p_{ss}^{n_{ss}}.$$

   Therefore

   $$\hat{p}_{dd} = \frac{n_{dd}}{n_{dd} + n_{ds} + 1}$$

   and

   $$\hat{p}_{ss} = \frac{n_{ss}}{n_{ss} + n_{ds}}.$$

2. $k = 2$:

In this case $F(p_{ss}) = 1$,

$$L = p_{dd}^{n_{dd}} x^{n_{ds}} p_{ss}^{n_{ss}} (1 - p_{dd} - x).$$

We solve this maximum problem more patiently. The log-likelihood is

$$l = n_{dd} \ln p_{dd} + n_{ds} \ln x + n_{ss} \ln p_{ss} + \ln (1 - p_{dd} - x).$$

All three parameters should be non-negative. The other constraint we have is

$$(1 - p_{dd})(1 - p_{ss}) - x \geq 0.$$

We use the multiplier v and construct the Lagrangian

$$l + v((1 - p_{dd})(1 - p_{ss}) - x).$$

The Kuhn-Tucker relations are

$$\begin{aligned}
\frac{n_{dd}}{p_{dd}} - \frac{1}{1 - p_{dd} - x} - v(1 - p_{ss}) &\leq 0,\ p_{dd}\left[\frac{n_{dd}}{p_{dd}} - \frac{1}{1 - p_{dd} - x} - v(1 - p_{ss})\right] = 0, \\
\frac{n_{ds}}{x} - \frac{1}{1 - p_{dd} - x} - v &\leq 0,\ x\left[\frac{n_{ds}}{x} - \frac{1}{1 - p_{dd} - x} - v\right] = 0, \\
\frac{n_{ss}}{p_{ss}} - v(1 - p_{dd}) &\leq 0,\ p_{ss}\left[\frac{n_{ss}}{p_{ss}} - v(1 - p_{dd})\right] = 0, \\
(1 - p_{dd})(1 - p_{ss}) - x &\geq 0,\ v[(1 - p_{dd})(1 - p_{ss}) - x] = 0, \\
v &\geq 0.
\end{aligned}$$

Assume that n_{dd}, n_{ds}, $n_{ss} \neq 0$, then p_{dd}, p_{ss}, $x \neq 0$, thus

$$\begin{aligned}
\frac{n_{dd}}{p_{dd}} - \frac{1}{1 - p_{dd} - x} - v(1 - p_{ss}) &= 0, \\
\frac{n_{ds}}{x} - \frac{1}{1 - p_{dd} - x} - v &= 0, \\
\frac{n_{ss}}{p_{ss}} - v(1 - p_{dd}) &= 0.
\end{aligned} \tag{8.20}$$

We have two cases:

Case 1: $v = 0$ and therefore $(1 - p_{dd})(1 - p_{ss}) - x \geq 0$. In this case the third equation in (8.20) cannot hold. So we need to consider the case when $v \neq 0$.

Case 2: $v \neq 0$ and thus $(1 - p_{dd})(1 - p_{ss}) - x = 0$. Then

$$\frac{1}{1 - p_{dd} - x} = \frac{1}{p_{ss}(1 - p_{dd})}$$

and from the third equation in (8.20) we have

$$v = \frac{n_{ss} + 1}{n_{ss} + n_{ds} + 1}. \tag{8.21}$$

By substitution of (8.21) into the second equation in (8.20) yields

$$\hat{p}_{ss} = \frac{n_{ss} + 1}{n_{ss} + n_{ds} + 1}$$

and the first equation yields

$$\hat{p}_{dd} = \frac{n_{dd}}{n_{dd} + n_{ds} + 1}.$$

This provides the solution for x as follows:

$$\hat{x} = \frac{(n_{ds} + 1)\, n_{ds}}{(n_{dd} + n_{ds} + 1)\,(n_{ss} + n_{ds} + 1)}.$$

Clearly, in this case

$$\hat{p}_{sd} = \frac{n_{ds}}{n_{ss} + n_{ds} + 1},$$

$$\hat{p}_{ds} = \frac{n_{ds} + 1}{n_{ss} + n_{ds} + 1}$$

and

$$\hat{p}_{sh} = \hat{p}_{dh} = 0.$$

3. $k = \infty$:

In this case

$$F(p_{ss}) = \frac{1}{1 - p_{ss}}.$$

The MLE equation for p_{ss} reduces to

$$\frac{p_{ss}}{1 - p_{ss}} = \frac{n_{ss}}{n_{ds}},$$

so

$$\hat{p}_{ss} = \frac{n_{ss}}{n_{ss} + n_{ds}} \tag{8.22}$$

and as always

$$\hat{p}_{dd} = \frac{n_{dd}}{n_{dd} + n_{ds} + 1}, \tag{8.23}$$

and

$$\hat{x} = \frac{n_{ds}\,(1 - \hat{p}_{ss})}{n_{dd} + n_{ds} + 1} = \frac{n_{ds}^2}{(n_{dd} + n_{ds} + 1)\,(n_{ss} + n_{ds})}.$$

From (8.22) and (8.23),

$$(1-\hat{p}_{dd})(1-\hat{p}_{ss}) = \frac{n_{ds}(n_{ds}+1)}{(n_{dd}+n_{ds}+1)(n_{ss}+n_{ds})} > \hat{x}.$$

Therefore all Kuhn-Tucker relations are satisfied. As regarding estimates of p_{ds} and p_{sd} they can be anything provided that

$$\begin{aligned} 0 &\leq \hat{p}_{ds} \leq \frac{n_{ds}+1}{n_{dd}+n_{ds}+1}, \\ 0 &\leq \hat{p}_{sd} \leq \frac{n_{ds}}{n_{ss}+n_{ds}}. \end{aligned}$$

and

$$\hat{p}_{ds}\hat{p}_{sd} = \frac{n_{ds}^2}{(n_{dd}+n_{ds}+1)(n_{ss}+n_{ds})}.$$

Therefore we have the extreme solution. When $\hat{p}_{ds}$ has a maximum value, we obtain the set of parameter estimates as follows

$$\begin{aligned} \hat{p}_{ds\,\max} &= \frac{n_{ds}+1}{n_{dd}+n_{ds}+1}, \\ \hat{p}_{sd\,\min} &= \frac{n_{ds}^2}{(n_{ds}+1)(n_{ss}+n_{ds})}, \\ \hat{p}_{dh\,\min} &= 0, \\ \hat{p}_{sh\,\max} &= \frac{n_{ds}}{(n_{ds}+1)(n_{ss}+n_{ds})}; \end{aligned}$$

whereas if $\hat{p}_{sd}$ has the maximum value, the set of parameter estimates is

$$\begin{aligned} \hat{p}_{sd\,\max} &= \frac{n_{ds}}{n_{ss}+n_{ds}}, \\ \hat{p}_{ds\,\min} &= \frac{n_{ds}}{n_{dd}+n_{ds}+1}, \\ \hat{p}_{dh\,\max} &= \frac{1}{n_{dd}+n_{ds}+1}, \\ \hat{p}_{sh\,\min} &= 0. \end{aligned}$$

8.1.2 The General Case Revisited

In general, we have the following Lagrange equations

$$\frac{n_{dd}}{p_{dd}} - \frac{1}{1-p_{dd}-x} - v(1-p_{dd}) = 0, \tag{8.24}$$

$$\frac{n_{ds}}{x} - \frac{F(p_{ss})}{1-p_{dd}-x} - v = 0, \tag{8.25}$$

$$\frac{n_{ss}}{p_{ss}} - \frac{xF'(p_{ss})}{1-p_{dd}-xF(p_{ss})} - v = 0, \tag{8.26}$$

$$\text{Constraint: } (1-p_{dd})(1-p_{ss})-x \geq 0, \tag{8.27}$$

$$(1-p_{dd})(1-p_{ss}) = x. \tag{8.28}$$

We already have obtained the solution of the first three equations (8.24), (8.25) and (8.26), when the Lagrange multiplier v is zero. We did not however examine the constraint, which we shall do now. The solution for p_{ss} was the root of the Fundamental Equation (8.18). From (8.14), the solution for x was

$$\hat{x} = \frac{n_{ds}}{(n_{dd}+n_{ds}+1)\,F(p_{ss})}. \tag{8.29}$$

From our constraint inequality (8.27),

$$\begin{aligned} x &\leq (1-p_{dd})(1-p_{ss}) \\ &\Leftrightarrow \frac{n_{ds}}{(n_{ds}+1)\,F(\hat{p}_{ss})} \leq 1-\hat{p}_{ss} \\ &\Leftrightarrow \frac{n_{ds}}{n_{ds}+1} \leq 1-\hat{p}_{ss}^{k-1}, \end{aligned}$$

and thus we need to have Jalali's *Fundamental Inequality*

$$\hat{p}_{ss}^{k-1} \leq \frac{1}{n_{ds}+1}. \tag{8.30}$$

The ML estimates of the parameters are as mentioned above. $\hat{p}_{ss}$ is the root of the Fundamental Equation (8.18) provided that this root satisfies the Fundamental Inequality (8.30). If the Fundamental Inequality is not satisfied by this root, then we need to solve all the four Lagrange equations and after doing so, we need to ensure that the multiplier v is not negative. It is easy to see that the solution for p_{dd}, using (8.24) and (8.25), is, as before, $\hat{p}_{dd} = \frac{n_{dd}}{n_{dd}+n_{ds}+1}$. The multiplier can then be found from (8.24) and (8.28) as follows

$$v = \frac{(n_{ds}+1)\,p_{ss}^{k-1}-1}{(1-p_{dd})(1-p_{ss})\,p_{ss}^{k-1}}. \tag{8.31}$$

(8.26) is then reduced to

$$\frac{n_{ss}(1-p_{ss})}{p_{ss}} - \frac{(1-p_{ss})^2 F'(p_{ss})}{p_{ss}^{k-1}} - \frac{p_{ss}^{k-1}(n_{ds}+1)-1}{p_{ss}^{k-1}} = 0. \tag{8.32}$$

But

$$(1-p_{ss})^2 F'(p_{ss}) = -(k-1)(1-p_{ss})\,p_{ss}^{k-2} + \left(1-p_{ss}^{k-1}\right). \tag{8.33}$$

Substituting (8.33) in (8.32) and simplifying, we get the solution for p_{ss} as follows

$$\hat{p}_{ss} = \frac{n_{ss}+k-1}{n_{ss}+n_{ds}+k-1}. \tag{8.34}$$

Note that, in this case the multiplier v should be non-negative. This means that

$$\hat{p}_{ss}^{k-1} \geq \frac{1}{n_{ds}+1}. \tag{8.35}$$

8.1.3 Summary

We can now sum up. Depending on whether the Fundamental Inequality holds or not we have two different expressions for the MLE of p_{dd}, p_{ds}, p_{dh}, p_{sd}, p_{sh} and p_{ss} (the MLE of p_{dd} always have the same expression).

If the Fundamental Inequality (8.30) holds, then $\hat{p}_{ss}$ is the root of the Fundamental Equation (8.18), and if the inequality holds strictly the estimates of at least one of the two parameters p_{dh} and p_{sh} is non-zero. As in the case of $k = \infty$, there is some degree of under-determination regarding parameters p_{ds}, p_{sd}, p_{dh} and p_{sh}. However, if one of these is chosen subject to the obvious constraints, then the rest can be determined uniquely.

If the Fundamental Inequality does not hold, then the following are the MLE of the parameters:

$$\begin{aligned}
\hat{p}_{dd} &= \frac{n_{dd}}{n_{dd}+n_{ds}+1}, \\
\hat{p}_{ss} &= \frac{n_{ss}+k-1}{n_{ss}+n_{ds}+k-1}, \\
\hat{p}_{ds} &= \frac{n_{ds}+1}{n_{dd}+n_{ds}+1}, \\
\hat{p}_{sd} &= \frac{n_{ds}}{n_{ss}+n_{ds}+k-1}, \\
\hat{p}_{dh} &= \hat{p}_{sh} = 0.
\end{aligned} \tag{8.36}$$

Remark 5 *The above dichotomy, as Jalali (2008c) puts it, is very intuitive. When k, the length of unknown E final subsequence is "short", there is nothing in the data to justify postulating state H, and therefore parameters p_{dh} and p_{sh} are estimated as zero. So state H will remain inaccessible from the other two states.*

If the length of unknown final subsequence is "long" though, we need to postulate state H as it is accessible at least from one of the states D and S.

We shall now illustrate how the set of estimates differs depending on whether the Fundamental Inequality holds or not. For example, when $k = 3$, the Fundamental Equation in (8.18) is reduced to

$$\frac{p_{ss}}{1+p_{ss}} = \frac{n_{ss}}{n_{ds}}.$$

The Fundamental Inequality (8.30), therefore, reduces to

$$n_{ss}^2 \leq n_{ds} - 2n_{ss}. \tag{8.37}$$

If (8.37) holds, then

$$\hat{p}_{ss} = \frac{n_{ss}}{n_{ds} - nlss},$$

and the Fundamental Inequality can ensure that

$$\hat{p}_{ss} \leq \frac{1}{1+n_{ss}}.$$

It also follows that

$$\hat{x} = \frac{n_{ds} - n_{ss}}{n_{dd} + n_{ds} + 1}.$$

If the Fundamental Inequality (8.30) does not hold, then

$$\hat{p}_{ss} = \frac{n_{ss} + 2}{n_{ds} + n_{ss} + 2}.$$

In this case, of course the estimated model becomes essentially a two-state model.

8.1.4 Independent Samples of Sequences

In this subsection we follow again Jalali (2008c) and consider the problem of estimation from a sample of sequences of size m. We assume that all sequences in the sample start at state D. For economising on subscript we denote, for the sample i the number of transitions from D to D by n_{0i}, from D to known S by n_{1i}, and from known S to known S by n_{2i}. We denote the average (over the sample) of these numbers, respectively by $\bar{n}_0$, $\bar{n}_1$ and $\bar{n}_2$. We denote by k_i the length of the unknown final subsequence of sequence i. Then the likelihood function for the whole sample is

$$\begin{aligned} L &= \prod_i p_{dd}^{n_{0i}} x^{n_{1i}} p_{ss}^{n_{2i}} \times \prod_{i=k_i \neq 0} \left(1 - p_{dd} - xF_i\left(p_{ss}\right)\right) \\ &= \left(p_{dd}^{\bar{n}_0} x^{\bar{n}_1} p_{ss}^{\bar{n}_2}\right)^m \prod_{i=k_i \neq 0} \left(1 - p_{dd} - xF_i\left(p_{ss}\right)\right), \end{aligned} \tag{8.38}$$

where $x = p_{ds}p_{sd}$ and $F_i\left(p_{ss}\right) = \dfrac{1 - p_{ss}^{k_i - 1}}{1 - p_{ss}}$. Upon finding the log-likelihood, finding its derivative with respect to the three parameters and equating them to zero, we shall obtain the following MLE equations

$$\frac{\bar{n}_0}{p_{dd}} - \frac{1}{m} \sum_{i=k_i \neq 0} \frac{1}{1 - p_{dd} - xF_i\left(p_{ss}\right)} = 0 \Rightarrow \bar{n}_0 = \frac{1}{m} \sum_{i=k_i \neq 0} \frac{p_{dd}}{1 - p_{dd} - xF_i\left(p_{ss}\right)}, \tag{8.39}$$

$$\frac{\bar{n}_1}{x} - \frac{1}{m} \sum_{i=k_i \neq 0} \frac{F_i\left(p_{ss}\right)}{1 - p_{dd} - xF_i\left(p_{ss}\right)} = 0 \Rightarrow \bar{n}_1 = \frac{1}{m} \sum_{i=k_i \neq 0} \frac{xF_i\left(p_{ss}\right)}{1 - p_{dd} - xF_i\left(p_{ss}\right)}, \tag{8.40}$$

$$\frac{\bar{n}_2}{p_{ss}} - \frac{1}{m} \sum_{i=k_i \neq 0} \frac{xF_i'\left(p_{ss}\right)}{1 - p_{dd} - xF_i\left(p_{ss}\right)} = 0 \Rightarrow \bar{n}_2 = \frac{1}{m} \sum_{i=k_i \neq 0} \frac{p_{ss}xF_i'\left(p_{ss}\right)}{1 - p_{dd} - xF_i\left(p_{ss}\right)}. \tag{8.41}$$

The constraint that we have, apart from the fact that p_{dd} and p_{ss} are probabilities is

$$\text{Constraint: } (1-p_{dd})(1-p_{ss}) \geq x.$$

If in the above, the equality holds, then state H becomes isolated from the rest, and we have essentially a two state Markov chain. From (8.39) and (8.40), it follows that

$$\bar{n}_0 + \bar{n}_1 + \rho = \frac{1}{m} \sum_{i=k_i \neq 0} \frac{1}{1 - p_{dd} - xF_i(p_{ss})} = \frac{\bar{n}_0}{p_{dd}} \tag{8.42}$$

where ρ is the proportion of those sequences in the sample which end with at least one unknown state. Hence

$$\hat{p}_{dd} = \frac{\bar{n}_0}{\bar{n}_0 + \bar{n}_1 + \rho}. \tag{8.43}$$

We note all terms like $\dfrac{xF_i(p_{ss})}{1 - p_{dd} - xF_i(p_{ss})}$ are increasing in x. So they achieve their maximum at $x = (1-p_{dd})(1-p_{ss})$. This maximum is

$$\frac{1 - p_{ss}^{k_i - 1}}{p_{ss}^{k_i - 1}} = p_{ss}^{-(k_i - 1)} - 1. \tag{8.44}$$

Hence, if (8.40) holds, then we should have

$$\frac{1}{m} \sum_{i=k_i \neq 0} \hat{p}_{ss}^{-(k_i - 1)} \geq \bar{n}_1 + \rho. \tag{8.45}$$

This is Jalali's new *Fundamental Inequality*, which, if it does not hold strictly, then our model essentially reduces to a two state model. According to whether this inequality hold or not, we have two kinds of solution as follows:

1. $\frac{1}{m} \sum_{i=k_i \neq 0} \hat{p}_{ss}^{-(k_i - 1)} \leq \bar{n}_1 + \rho$

   In this case (8.40) cannot hold (unless the equality holds, which means that $x = (1-p_{dd})(1-p_{ss})$ and therefore we obtain an essentially two state model). Then it follows from Kuhn-Tucker style argument that we need once again to have the constraint equality satisfied. Hence the likelihood function is of the form

$$\begin{aligned} L &= \left(p_{dd}^{\bar{n}_0} x^{\bar{n}_1} p_{ss}^{\bar{n}_2}\right)^m \prod_{i=k_i \neq 0} (1 - p_{dd} - xF_i(p_{ss})) \\ &= \left(\left(p_{dd}^{\bar{n}_0}! - p_{dd}\right)^{\bar{n}_1} (1 - p_{ss})^{\bar{n}_1} p_{ss}^{\bar{n}_2}\right)^m (1 - p_{dd})^{\rho m} \prod_{i=k_i \neq 0} p_{ss}^{k_i - 1}. \end{aligned} \tag{8.46}$$

   This is maximised at

$$\hat{p}_{dd} = \frac{\bar{n}_0}{\bar{n}_0 + \bar{n}_1 + \rho} \tag{8.47}$$

and

$$\hat{p}_{ss} = \frac{\bar{n}_2 - \rho + \frac{1}{m}\sum_{i=k_i\neq 0} k_i}{\bar{n}_1 + \bar{n}_2 - \rho + \frac{1}{m}\sum_{i=k_i\neq 0} k_i}. \tag{8.48}$$

As $\hat{p}_{ds} = 1 - \hat{p}_{dd}$, $\hat{p}_{sd} = 1 - \hat{p}_{ss}$ and $\hat{p}_{dh} = \hat{p}_{sh} = 0$.

2. $\frac{1}{m}\sum_{i=k_i\neq 0} \hat{p}_{ss}^{-(k_i-1)} \geq \bar{n}_1 + \rho$

In this case the MLE equations can be solved and the solution satisfies the constraint inequality strictly. We already have obtained p_{dd}. The solution for p_{ss} and x can be obtained from the second and third MLE equations when we let $\hat{p}_{dd}$ to equal $\frac{\bar{n}_0}{\bar{n}_0 + \bar{n}_1 + \rho}$. As in the case $m = 1$, we cannot obtain unique values for p_{ds}, p_{sd}, p_{dh} and p_{sh} but choosing any of these subject to the constraint, the rest can be found uniquely.

8.1.5 Important Practical Case

In such cases that the length of the sequences can be controlled by the experimenter, to ease the solution of equations, the experimenter may fix k for all sequences. This means that in the case of each sequence the experiment continues until k consecutive states E appear and then it stops. This is rather similar to Type II censoring. We, therefore, have, for all i, $k_i = k$. In this case it is reasonable to assume that the common k is not zero.

The likelihood function then will be

$$L = \left(p_{dd}^{\bar{n}_0} x^{\bar{n}_1} p_{ss}^{\bar{n}_2} \left(1 - p_{dd} - xF\left(p_{ss}\right)\right)\right)^m. \tag{8.49}$$

It is obvious that the new maximisation problem is identical to the single sequence case. By replacing n_{dd} as $\bar{n}_0$, n_{ds} as $\bar{n}_1$ and n_{ss} as $\bar{n}_2$, $\hat{a}$ is the root of the Fundamental Equation (8.18) whereas x is given by (8.29). This ends the exposition of Jalali (2008c).

8.2 Simulation Studies

To investigate the performance of SRAMPT in estimating the parameters of the epidemiology model, we simulated a set of Markov chains using Matlab by considering different scenarios for the epidemic chain. For each scenario, we simulated five samples to study, each consisting of twenty independent sequences. We terminated the simulation of each sequence after observing five consecutive E's.

8.2.1 Scenario 1: A Typical Example

In this scenario, we assume that the probability that a contact of a diseased host to be sub-clinically infected is half of p_{dd}, the probability that a contact of a sub-clinically infected host to show disease is one third of p_{dd} while the probability that the contact is sub-clinically

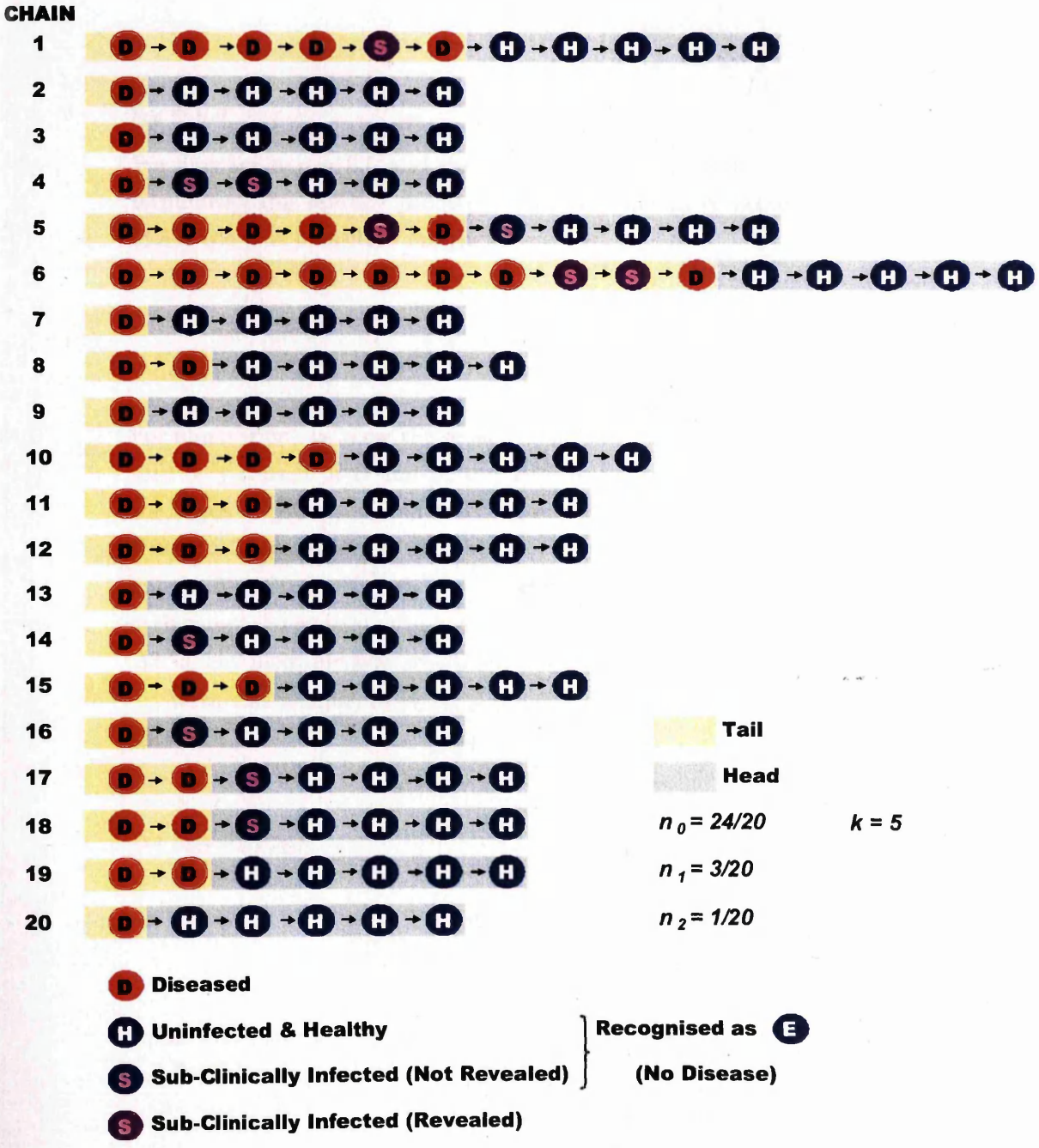


Figure 8.2: An illustration of a sample consisting of twenty simulated Markov chains ($p_{dd} = 0.5$) for the Epidemiology Model.

infected is one sixth of p_{dd}. Therefore, the transition probabilities in (8.1) are

$$\boldsymbol{P} = \begin{bmatrix} p_{dd} & 1 - \frac{3p_{dd}}{2} & \frac{p_{dd}}{2} \\ 0 & 1 & 0 \\ \frac{p_{dd}}{3} & 1 - \frac{p_{dd}}{2} & \frac{p_{dd}}{6} \end{bmatrix}. \tag{8.50}$$

$p_{dd} = 0.5$

We first set $p_{dd} = 0.5$ and simulated five samples, each with 20 sequences in which we terminated the Markov process when five consecutive states of E appeared. Following (8.50) the true transition matrix is

$$\boldsymbol{P} = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0 & 1 & 0 \\ 0.1667 & 0.75 & 0.0833 \end{bmatrix}.$$

For illustration, we show the sequences of Sample 4 in Figure 8.2. As seen in the figure, the tails only consist of D and S, and all S in the tails are self-revealed. On the other hand, the heads are all recognised as E, since all S in the heads are not revealed. In this example there are 24 transitions from D to D, 3 transitions from D to known S and 1 transition from known S to S; $\bar{n}_o = \frac{24}{20}$, $\bar{n}_1 = \frac{3}{20}$ and $\bar{n}_2 = \frac{1}{20}$. We then applied the SRAMPT method to each sample to estimate the transition probabilities. The results are presented in Table 8.2. We hereby illustrate how we obtain the estimates for Sample 4.

With this sample, $m = 20$, $\rho = 1$, $\bar{n}_o = \frac{24}{20}$, $\bar{n}_1 = \frac{3}{20}$ and $\bar{n}_2 = \frac{1}{20}$. The likelihood is therefore

$$L = \left(p_{dd}^{\frac{24}{20}} x^{\frac{3}{20}} p_{ss}^{\frac{1}{20}} \left(1 - p_{dd} - xF\left(p_{ss}\right)\right) \right)^{20}.$$

We first find $\hat{p}_{dd}$:

$$\hat{p}_{dd} = \frac{\bar{n}_0}{\bar{n}_0 + \bar{n}_1 + \rho} = \frac{\frac{24}{20}}{\frac{24}{20} + \frac{3}{20} + 1} = \frac{24}{47} = 0.5106.$$

Upon substituting k, $\bar{n}_1$ and $\bar{n}_2$ into (8.18), we solve the Fundamental Equation (8.18) for $\hat{p}_{ss}$, for this sample the estimate is

$$\hat{p}_{ss} = 0.2602.$$

With $\hat{p}_{dd}$, $\hat{p}_{ss}$, $\bar{n}_1$ and k, we shall now find $\hat{x}$ from (8.29), where

$$\hat{x} = \frac{\frac{3}{20}\left(\frac{23}{47}\right)}{F(0.2602)\left(\frac{3}{20}+1\right)} = 0.0474.$$

Now, we find the range of values of the remaining four parameters p_{ds}, p_{sd}, p_{dh} and p_{sh}. We find the maximum value of $\hat{p}_{ds}$ by setting $\hat{p}_{dh}$ as 0, so

$$\hat{p}_{ds\,\max} = 1 - \hat{p}_{dd} = \frac{23}{47} = 0.4894$$

and since $\hat{x} = 0.0474$, the minimum value of $\hat{p}_{sd}$ is given by

$$\hat{p}_{sd\,\min} = \frac{\hat{x}}{\hat{p}_{ds\,\max}} = \frac{0.0474}{0.4894} = 0.0969$$

whereas the maximum value is obtained by setting $\hat{p}_{sh}$ as 0, so

$$\hat{p}_{sd\,\max} = 1 - \hat{p}_{ss} = 0.7398.$$

With $\hat{p}_{sd\,\max}$, the minimum value of $\hat{p}_{ds}$ is then

$$\hat{p}_{ds\,\min} = \frac{\hat{x}}{\hat{p}_{sd\,\max}} = \frac{0.0474}{0.7398} = 0.0641.$$

We know the minimum values of $\hat{p}_{dh}$ and $\hat{p}_{sh}$ are both 0, and their maximum values are as follows

$$\begin{aligned}
\hat{p}_{dh\,\max} &= 1 - \hat{p}_{dd} - \hat{p}_{ds\,\min} = 1 - \frac{24}{47} - 0.0641 = 0.4253, \\
\hat{p}_{sh\,\max} &= 1 - \hat{p}_{ss} - \hat{p}_{sd\,\min} = 1 - 0.2602 - 0.0969 = 0.6429.
\end{aligned}$$

We shall now summarise the estimates of all parameters:

$$\begin{aligned}
\hat{p}_{dd} &= 0.5106, \\
\hat{p}_{ss} &= 0.2602, \\
0 \leq\ & \hat{p}_{dh} \leq 0.4253, \\
0.0641 \leq\ & \hat{p}_{ds} \leq 0.4894, \\
0.0969 \leq\ & \hat{p}_{sd} \leq 0.7398, \\
0 \leq\ & \hat{p}_{sh} \leq 0.6429.
\end{aligned}$$

Table 8.1 presents $\bar{n}_0$, $\bar{n}_1$ and $\bar{n}_2$ for each of the five samples. Since the true value of p_{ss} is only 0.0833, we did not obtain any transition from S to S in the tail of samples 1, 3 and 5. This is why the estimates of p_{ss} for these samples are 0, as seen in Table 8.2. The true

Sample	$\bar{n}_0$	$\bar{n}_1$	$\bar{n}_2$
1	$\frac{24}{20}$	$\frac{1}{20}$	0
2	$\frac{17}{20}$	$\frac{4}{20}$	$\frac{1}{20}$
3	$\frac{18}{20}$	$\frac{2}{20}$	0
4	$\frac{24}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
5	$\frac{27}{20}$	$\frac{1}{20}$	0

Table 8.1: Details of transitions in each sample for Scenario 1 with $p_{dd} = 0.5$.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.5333	0	(0.0222, 0.4667)	(0.0476, 1)	(0, 0.4444)	(0, 0.9524)	0.0222
2	0.4146	0.2045	(0.0977, 0.5854)	(0.1328, 0.7955)	(0, 0.4877)	(0, 0.6627)	0.0777
3	0.4500	0	(0.0500, 0.5500)	(0.0909, 1)	(0, 0.5000)	(0, 0.9091)	0.0500
4	0.5106	0.2602	(0.0641, 0.4894)	(0.0969, 0.7398)	(0, 0.4253)	(0, 0.6429)	0.0474
5	0.5625	0	(0.0208, 0.4375)	(0.0476, 1)	(0, 0.4167)	(0, 0.9524)	0.0208
True Value	0.5000	0.0833	0.2500	0.1667	0.2500	0.7500	0.0417

Table 8.2: SRAMPT estimates of transition probabilities for Scenario 1 with $p_{dd} = 0.5$.

values of all four parameters p_{ds}, p_{sd}, p_{dh} and p_{sh} do lie within the range of the estimates provided by SRAMPT. The estimate of p_{dd} is satisfactory for each sample, close to the true value 0.5. The Fundamental Inequality (8.45) holds in each sample, so it remains as a three-state Markov chain in each case.

Since we only have a range of values for $\hat{p}_{dh}$ and $\hat{p}_{sh}$, and they are dependant on each other with the following relationship:

$$p_{sh} = 1 - p_{ss} - \frac{x}{1 - p_{dd} - p_{dh}}, \tag{8.51}$$

we plot $\hat{p}_{sh}$ versus $\hat{p}_{dh}$, alongside a 45° line, for each sample in Figure 8.3 to understand how likely is $\hat{p}_{sh}$ greater than $\hat{p}_{dh}$. From these plots we observe that $\hat{p}_{sh}$ is more likely to be greater than $\hat{p}_{dh}$, which is true because p_{sh} is 0.75 and p_{dh} is 0.25.

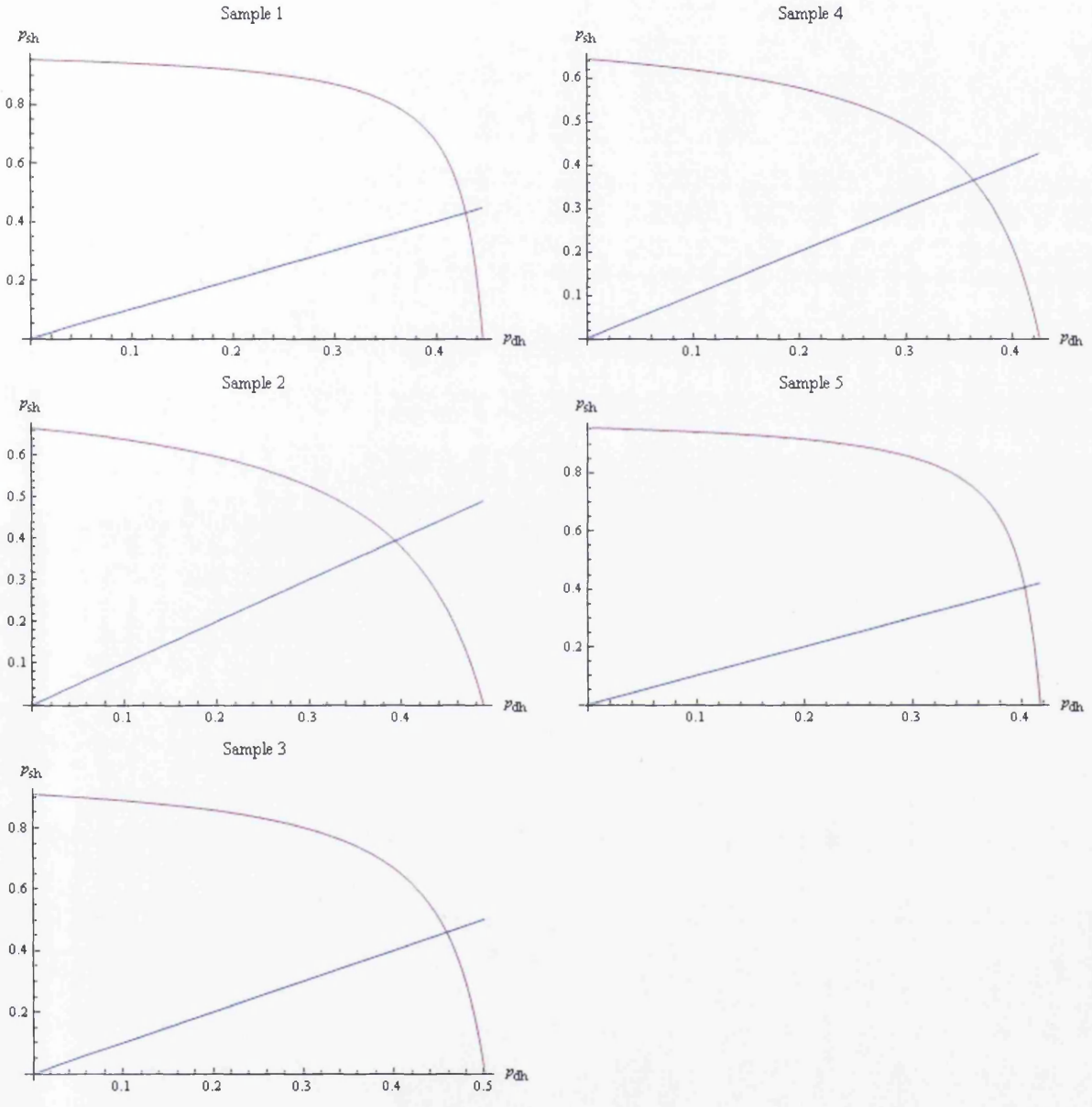


Figure 8.3: Plots of $\hat{p}_{sh}$ versus $\hat{p}_{dh}$ for Scenario 1 with $p_{dd} = 0.5$.

Inclusion of Domain Knowledge

Table 8.2 shows that we can obtain reasonable estimates for some of the parameters (p_{dd}, p_{ss} and x) but for the others, our estimates vary over a considerable range. We now consider how assumptions based on the specific epidemiological system or disease being studied (domain knowledge) might help us narrow down the estimates within these ranges. We are interested in whether the assumptions must be strong, for example where we might know the value of one of the parameters *a priori*, or weak, where we might only know the approximate relationship between parameters. We start with a weak assumption, and one that is likely to apply to most disease systems. Although we are unlikely to know the value of the parameters p_{sh} and p_{dh}, in general we can make the assumption that $p_{sh} > p_{dh}$ and they are not likely to be similar in magnitude. This is simply the assumption that sub-clinical infections are less likely to transmit infection than clinical disease infections. The shape of the curves in Figure 8.3 is helpful here. In our example, we assume that p_{sh} should lie between about 0.5 and 0.6429 (otherwise it is close in value of p_{dh}), and therefore estimate it as the mid-point of these two values, which is

$$\hat{p}_{sh} = \frac{0.5 + 0.6429}{2} = 0.5715.$$

Note that we cannot now directly estimate p_{dh}, since there is a considerable range of values consistent with our key assumption. Instead, first we find $\hat{p}_{sd}$ as follows

$$\hat{p}_{sd} = 1 - \hat{p}_{ss} - \hat{p}_{sh} = 0.1684,$$

and $\hat{p}_{ds}$ can now be obtained as

$$\hat{p}_{ds} = \frac{\hat{x}}{\hat{p}_{sd}} = 0.2818.$$

$\hat{p}_{dh}$ is then given by

$$\hat{p}_{dh} = 1 - \hat{p}_{dd} - \hat{p}_{ds} = 0.2076.$$

Thus, with a weak assumption for the relative magnitude of p_{sh} and p_{dh}, we managed to find quite accurate estimates of all parameters. Based on the assumption, the estimates of these parameters for each sample are summarised in Table 8.3. We can see that all estimates are now quite close to the true values.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.5333	0	0.0993	0.2238	0.3674	0.7762	0.0222
2	0.4146	0.2045	0.3630	0.2142	0.2224	0.5814	0.0777
3	0.4500	0	0.2037	0.2455	0.3463	0.7546	0.0500
4	0.5106	0.2602	0.2818	0.1684	0.2076	0.572	0.0474
5	0.5625	0	0.0931	0.2238	0.3444	0.7762	0.0208
True Value	0.5	0.0833	0.25	0.1667	0.25	0.75	0.0417

Table 8.3: SRAMPT estimates of transition probabilities based on a weak assumption for Scenario 1 with $p_{dd} = 0.5$.

$p_{dd} = 0.2$

Next we investigate a similarly structured example, but where disease transmission is less likely. We set p_{dd} as 0.2, the rest of the parameters are set according to (8.50), and hence

$$\boldsymbol{P} = \begin{bmatrix} 0.2 & 0.7 & 0.1 \\ 0 & 1 & 0 \\ 0.0667 & 0.9 & 0.0333 \end{bmatrix}.$$

In this case, the chance for a diseased host to pass disease to a contact is low, so is the chance for the contact to be sub-clinically infected. In this scenario we initially simulated samples of 20 chains, however due to the rare transfer of disease to sub-clinical (and even rarer transfer from sub-clinical to sub-clinical) we found that this sample size was far too low to enable estimates of the parameters, other than p_{dd}. Instead, we simulated 50 chains per sample. The details of the transitions in these five samples are summarised in Table 8.4. With a larger number of sequences, we observed at least one transition from D to S in every sample. Table 8.5 shows that with this sample size, we obtained good estimates of p_{dd} and x. Since we did not observe any transition from S to S in four samples, $\hat{p}_{ss}$ is estimated as zero. Indeed, the true value of p_{ss} is 0.0333 which means that we will only observe three transitions from S to S in 100 transitions from S. Since each sequence was started with D, and the true value of p_{ds} is 0.1, it is very unlikely to observe an S to S transition. It is worth noting that Sample 1 is the only sample where we observe one transition from S to S, however in this case $\hat{p}_{ss}$ has been severely over-estimated. According to the relationship (8.51), we plot $\hat{p}_{sh}$ against $\hat{p}_{dh}$ in Figure 8.4. We can see that the estimated relationship between the two parameters is again very helpful. Apart from Sample 1, it is clear that $\hat{p}_{sh}$ is likely to be greater than 0.8 and larger than $\hat{p}_{dh}$. So we estimate $\hat{p}_{sh}$ as the mid-point of 0.8 and the maximum value of $\hat{p}_{sh}$. With an exact estimate of $\hat{p}_{sh}$, we can now estimate $\hat{p}_{sd}$, $\hat{p}_{ds}$ and $\hat{p}_{dh}$. The results are expressed in Table 8.6. Again, even with a weak domain assumption, the estimates of all parameters, except from Sample 1, are close to the true values.

Sample	$\bar{n}_0$	$\bar{n}_1$	$\bar{n}_2$
1	$\frac{12}{50}$	$\frac{1}{50}$	$\frac{1}{50}$
2	$\frac{10}{50}$	$\frac{2}{50}$	0
3	$\frac{13}{50}$	$\frac{1}{50}$	0
4	$\frac{18}{50}$	$\frac{1}{50}$	0
5	$\frac{15}{50}$	$\frac{2}{50}$	0

Table 8.4: Details of transitions in each sample for Scenario 1 with $p_{dd} = 0.2$.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.1905	0.6573	(0.0195, 0.8095)	(0.0083, 0.3427)	(0, 0.7900)	(0, 0.3344)	0.0067
2	0.1613	0	(0.0323, 0.8387)	(0.0385, 1)	(0, 0.8065)	(0, 0.9615)	0.0323
3	0.2032	0	(0.0196, 0.7969)	(0.0196, 1)	(0, 0.7813)	(0, 0.9804)	0.0156
4	0.2609	0	(0.0145, 0.7391)	(0.0196, 1)	(0, 0.7246)	(0, 0.9804)	0.0145
5	0.2239	0	(0.0299, 0.7761)	(0.0385, 1)	(0, 0.7462)	(0, 0.9615)	0.0299
True Value	0.2	0.0333	0.1	0.0667	0.7	0.9	0.0067

Table 8.5: SRAMPT estimates of transition probabilities for Scenario 1 with $p_{dd} = 0.2$.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.1905	0.6573	0.0098	0.0255	0.7997	0.3172	0.0067
2	0.1613	0	0.2706	0.1192	0.5682	0.8808	0.0323
3	0.2032	0	0.1423	0.1098	0.6546	0.8902	0.0156
4	0.2609	0	0.1320	0.1098	0.6071	0.8902	0.0145
5	0.2239	0	0.2504	0.1192	0.5258	0.8808	0.0299
True Value	0.2	0.0333	0.1	0.0667	0.7	0.9	0.0067

Table 8.6: SRAMPT estimates of transition probabilities based on a weak assumption for Scenario 1 with $p_{dd} = 0.2$.

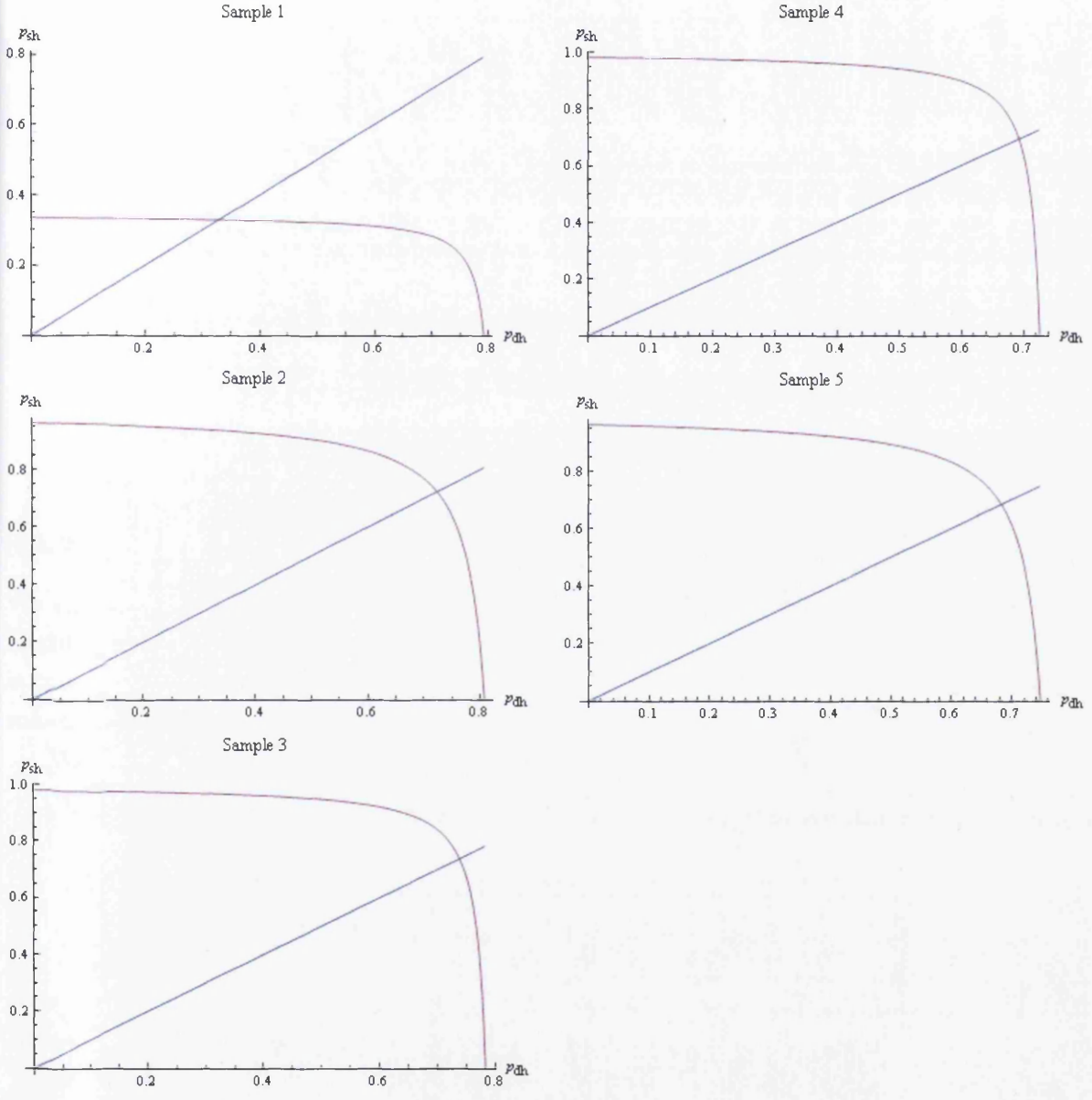


Figure 8.4: Plots of $\hat{p}_{sh}$ versus $\hat{p}_{dh}$ for Scenario 1 with $p_{dd} = 0.2$.

Sample	$\bar{n}_0$	$\bar{n}_1$	$\bar{n}_2$
1	$\frac{25}{20}$	$\frac{8}{20}$	$\frac{3}{20}$
2	$\frac{19}{20}$	$\frac{8}{20}$	$\frac{2}{20}$
3	$\frac{21}{20}$	$\frac{6}{20}$	$\frac{1}{20}$
4	$\frac{14}{20}$	$\frac{10}{20}$	$\frac{4}{20}$
5	$\frac{24}{20}$	$\frac{8}{20}$	$\frac{4}{20}$

Table 8.7: Details of transitions in each sample for Scenario 2 with $p_{dd} = 0.4$.

8.2.2 Scenario 2: Strong Sub-clinical Effects

We now consider a scenario that is less likely to be found in epidemiology, but one which might have different consequences for parameter estimation. Here, the chance that disease is passed on is assumed to be the same regardless of whether the infection is clinical or sub-clinical. In our first example, we use high probabilities for disease transfer.

We set the transition probabilities as follows:

$$\boldsymbol{P} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0 & 1 & 0 \\ 0.4 & 0.3 & 0.3 \end{bmatrix} \tag{8.52}$$

i.e. a host in state D or S has the same behaviour: there is a 40% chance to transmit disease to a contact, 30% chance to cause the contact to be sub-clinically infected and 30% chance that the contact will be uninfected.

In Table 8.7 we see that our simulated sample have more revealed states S and this makes our data richer; we observe more transitions from S to D and S to S. Thus the ranges of p_{ds}, p_{sd}, p_{dh} and p_{sh} are smaller (Table 8.8).

If we are now to consider how knowledge of the specific epidemiology may refine the parameter estimates we must make a different assumption to our first scenario. Here the model has been chosen to reflect a situation of strong sub-clinical effects, so our domain assumption is $p_{sd} = p_{dd}$. The updated estimates are given in Table 8.9 which shows good agreement with the true values.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.4717	0.2869	(0.1520, 0.5283)	(0.2051, 0.7131)	(0, 0.3763)	(0, 0.5080)	0.1084
2	0.4043	0.2045	(0.1705, 0.5957)	(0.2277, 0.7955)	(0, 0.4252)	(0, 0.5678)	0.1356
3	0.4468	0.1441	(0.1277, 0.5532)	(0.2605, 0.8559)	(0, 0.4255)	(0, 0.6583)	0.1093
4	0.3182	0.3026	(0.2292, 0.6818)	(0.2344, 0.6974)	(0, 0.4526)	(0, 0.4630)	0.1599
5	0.4615	0.3635	(0.1566, 0.5385)	(0.1851, 0.6365)	(0, 0.3819)	(0, 0.4514)	0.0997
True Value	0.4	0.3	0.3	0.4	0.3	0.3	0.12

Table 8.8: SRAMPT estimates of transition probabilities for Scenario 2 with $p_{dd} = 0.4$.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.4717	0.2869	0.2298	0.4717	0.2985	0.2414	0.1084
2	0.4043	0.2045	0.3354	0.4043	0.2603	0.3912	0.1356
3	0.4468	0.1441	0.2446	0.4468	0.3086	0.4091	0.1093
4	0.3182	0.3026	0.5025	0.3182	0.1793	0.3792	0.1599
5	0.4615	0.3635	0.2160	0.4615	0.3225	0.1750	0.0997
True Value	0.4	0.3	0.3	0.4	0.3	0.3	0.12

Table 8.9: SRAMPT estimates of transition probabilities based on a strong assumption for Scenario 2 with $p_{dd} = 0.4$.

We lastly investigate a similar situation, but for lower rates of disease transmission. We consider the transition matrix

$$\boldsymbol{P} = \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0 & 1 & 0 \\ 0.1 & 0.6 & 0.3 \end{bmatrix}$$

It is less likely (only 10% chance) for a diseased host to transmit disease. Most of the contacts (60%) of a diseased host are uninfected and healthy, whereas there is a 30% chance that the contact is sub-clinically infected. The contacts of a sub-clinically infected host has exactly the same probabilities to state D, H and S like the diseased host.

The simulated data sets are given in Table 8.10 and the parameter estimates in Table 8.10. The ranges are, as expected, higher. However, with the inclusion of the assumption $p_{sd} = p_{dd}$, a good set of parameter estimates are obtained in 3 samples out of 5. The exceptions are Samples 1 and 4. In Sample 1, we observed an equal number of transition from S to S and D to S, hence p_{ss} has been over-estimated (about twice the true value) and this affected the estimation of the other parameters.

It is worth mentioning that we treated Sample 4 differently. Since the lower bound of $\hat{p}_{sd}$ (= 0.0566) is larger than $\hat{p}_{dd}$, we assume that $\hat{p}_{sd}$ is 0.0566, instead of 0.0536. Since $\hat{x}$ is close to $\hat{p}_{sd}$, $\hat{p}_{ds}$ is over-estimated, and hence $\hat{p}_{dh}$ is zero.

Sample	$\bar{n}_0$	$\bar{n}_1$	$\bar{n}_2$
1	$\frac{9}{50}$	$\frac{1}{50}$	$\frac{1}{50}$
2	$\frac{9}{50}$	$\frac{3}{50}$	$\frac{2}{50}$
3	$\frac{9}{50}$	$\frac{2}{50}$	$\frac{1}{50}$
4	$\frac{3}{50}$	$\frac{3}{50}$	0
5	$\frac{4}{50}$	$\frac{3}{50}$	$\frac{2}{50}$

Table 8.10: Details of transitions in each sample for Scenario 2 with $p_{dd} = 0.1$.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.1500	0.6573	(0.0205, 0.8500)	(0.0083, 0.3427)	(0, 0.8295)	(0, 0.3344)	0.0070
2	0.1452	0.4613	(0.0507, 0.8548)	(0.0319, 0.5387)	(0, 0.8042)	(0, 0.5067)	0.0273
3	0.1475	0.3635	(0.0334, 0.8525)	(0.0249, 0.6365)	(0, 0.8191)	(0, 0.6116)	0.0212
4	0.0536	0	(0.0536, 0.9464)	(0.0566, 1)	(0, 0.8929)	(0, 0.9434)	0.0536
5	0.0702	0.4613	(0.0551, 0.9298)	(0.0319, 0.5387)	(0, 0.8747)	(0, 0.5067)	0.0297
True Value	0.1	0.3	0.3	0.1	0.6	0.6	0.03

Table 8.11: SRAMPT estimates of transition probabilities for Scenario 2 with $p_{dd} = 0.1$.

Sample	$\hat{p}_{dd}$	$\hat{p}_{ss}$	$\hat{p}_{ds}$	$\hat{p}_{sd}$	$\hat{p}_{dh}$	$\hat{p}_{sh}$	$\hat{x}$
1	0.1500	0.6573	0.0467	0.1500	0.8033	0.1927	0.007
2	0.1452	0.4613	0.1880	0.1452	0.6668	0.3935	0.0273
3	0.1475	0.3635	0.1437	0.1475	0.7088	0.489	0.0212
4	0.0536	0	0.9464	0.0566	0	0.9434	0.0536
5	0.0702	0.4613	0.4231	0.0702	0.5067	0.4685	0.0297
True Value	0.1	0.3	0.3	0.1	0.6	0.6	0.03

Table 8.12: SRAMPT estimates of transition probabilities based on a weak assumption for Scenario 2 with $p_{dd} = 0.1$.

8.3 Discussion and Summary

In this chapter we propose that the model for transmission of sub-clinical infection, developed for the study of scrapie in Chapter 7, has applications for modelling chains of infection in an epidemiological setting. Instead of experimental transfer of infection, we now consider a chain of contacts. The only assumption that must be relaxed is that D is now no longer an absorbing state, i.e. contact with a clinical case does not necessarily result in transfer of infection.

In the first part of this chapter we show that the additional parameters in the transition matrix of the Markov model greatly complicate the estimation procedure. Jalali (2008c) studied the problem by looking at a chain of states. Using the SRAMPT model, he showed how the estimation problem can be solved using the MLE. We first discuss how his model can be used to solve the estimation problem. Consider a chain of states, which can be divided into two parts: head and tail. The tail of a chain is a sequence of D's and S's, ending with a D; whereas the head, which follows the tail, only contains E (we cannot tell if a node is healthy or sub-clinically infected in the head). Since in the tail of a chain, $n_{ds} = n_{sd}$, x was set to be the product of p_{ds} and p_{sd}. With SRAMPT, we can only obtain the exact values of p_{dd}, p_{ss} and x. With these three parameter estimates, we can then find the extreme solutions of p_{ds}, p_{sd}, p_{dh} and p_{sh}. We have two different expressions for the parameters, depending on whether the Fundamental Inequality in (8.45) holds or not. If it holds, p_{ss} is the root of the Fundamental Equation in (8.18) and the rest of the parameters can be obtained with $\hat{p}_{ss}$. Else, the parameters can be obtained by solving all the Lagrange equations, and the exact formulae of the parameters have been provided by Jalali (2008c), as shown in (8.36). In the latter case, the Markov model is reduced to two-state model, as H will remain inaccessible from the other two states.

In the second part of this chapter we illustrate the parameter estimation process with a set of examples, using simulated data. First we consider a general case where most disease transmission (whether to clinical status or sub-clinical status) occurs from state D. The parameter estimates are promising, in particular if knowledge can be applied from the systems under study to narrow down the ranges found from the raw data. In the examples, if we assume that $p_{sh} > p_{dh}$, then our parameter estimation is precise (although the sample size may need to be high, greater than 50 epidemic chains, for low overall disease transmission rates). For most diseases, this assumption should hold. For example, it is generally thought that contacts with clinical cases of tuberculosis (a bacterial infection) are an order or magnitude more infectious than asymptomatic carriers of the disease. For the measles virus, sub-clinical infection is known to occur in vaccinated individuals. Since measles epidemics may be sustained in vaccinated areas (Whittle *et al.* (1999)), p_{sh} may be closer in value to p_{dh} than for for tuberculosis, but the inequality should still clearly hold.

In our final examples, we consider a case where sub-clinical infection is highly transmissible. Due to the larger numbers of observed transitions between all states in the model,

we found that parameter estimation was somewhat easier in this case. We note that, even with no assumed domain knowledge, the ranges for the parameter estimates were smaller, and it was interesting that the mid-point of the range would provide a good point estimate. Although this scenario is less common in epidemiology, there are some potential application systems such as the common bacteria infection *Escherichia Coli*, where asymptomatic individuals are known to be infective.

Chapter 9

Summary and Future Directions of Research

In this chapter we provide an overview of our work in previous chapters, present our conclusions and propose areas for future study. Let us begin with a summary of each of the important topics studied in this thesis.

9.1 Summary and Conclusions

9.1.1 Mixtures of Exponential Distributions

The problem of estimating the parameters in an exponential mixture distribution has been discussed in depth in Chapter 3. Several methods which have formal similarities to the traditional method of moments and been devised by Jalali have been investigated, in particular for a two-component exponential mixture. The performances of these methods in fitting data sets simulated from an exponential mixture distribution have been compared with the more well known maximum likelihood approach.

The standard method of moments is obviously not favoured due to its poor performance in estimating the parameters, even when the sample size is large enough. Since the estimates are obtained by solving a quadratic equation, the roots are likely to be negative or in complex forms. Our investigation showed that one of the reasons for the implausibility of estimates is the high discrepancy between the raw moments and the theoretical moments. We have shown that the variance of the k^{th} moment $Var\left[\hat{\mu}_k\right]$ increases with k. We also noticed that when there is large separation between the two components, the effect of the smaller moment (from the second component) pales into insignificance on the overall moment. This leads to a poor estimate of the second rate parameter b, corresponding to the exponential with lower mean. Therefore, we need a method which controls the variation between the moments from the components.

By replacing the integer k with a fraction κ in the moment estimator, we observed great improvement with the method called the fractional moment estimator. Using a Tay-

lor expansion argument, we found the approximated values of the asymptotic variance of fractional moment estimator. The theoretical minimum variance of the estimator and its corresponding κ have sound agreement with the simulation results. Therefore, we are able to suggest the optimal κ for estimating a mixture of two exponential distributions with different separation: for mixtures with slightly separated components, we should use a large fraction; the ideal fraction decreases with the magnitude of the separation. In practice, we do not know the separation between the components. We suggested a way for users to choose an appropriate κ so that the precision of the estimates are guaranteed. One should simply fit a raw sample with a few different κ, and substitute the resulted sets of estimates into the formulae of the theoretical asymptotic variances of the estimators. The set of parameters which produces the minimum variance should be chosen.

Extending the method of fractional moments, we further added an attenuation $\exp(-cT)$, where c is the attenuation factor and T is an observation from a random sample, to the fractional moments. This modified method is named as the method of attenuated moments. Our simulation experiments showed that the attenuated moment estimation is an outstanding method which provides estimates which have small bias and have small variances, especially when the components are well separated, the sample size is large enough, or both. We successfully obtained a good approximation of the asymptotic variance of the attenuated moment estimator, which allows us to suggest the best combination of fraction κ and attenuation c for a two-component exponential mixture with different degrees of separation. As the components depart further from each other, we should use a lower fraction κ; whereas c should always be small. Like the fractional estimator, users should estimate the parameters with a few combinations of κ and c in real life, when the separation between the components is unknown. The set of parameters, when substituted into $\boldsymbol{V}\left[\hat{\boldsymbol{\Theta}}\right]$, which gives the lowest theoretical asymptotic variances of the estimators should be chosen.

The method of Appell moments is a modified moment-based method which makes use of Appell-Fourier sequences $\{h_k\}$, where h_k is an Appell function. We illustrated how Appell moments, which are particularly based on sequences of trigonometric functions, can be used to solve the estimation problem for a mixture of two exponential distributions. With $h_\alpha(t) = \sin^\alpha \omega t$, where α is the highest index, the α^{th} observed Appell moments are given by $n_o^{-1}\sum_{i=1}^{n_o} h_\alpha(t_i)$; we need $\alpha+1$ moments where the lower indexed $h_k(t)$ can be found by differentiation. Our simulations considered $\alpha = 3$, 4 and 5; the estimation results are satisfactory only when $\alpha = 3$. Since the value of ω has a significant effect on the precision of the estimates, we evaluated the approximated asymptotic variance of Appell moment estimators for $\alpha = 3$ and used it to find the optimal value of ω which provides estimates with the minimum variances. Apparently, the optimal ω increases as the components are better separated.

The method using order statistics, again devised by Jalali, is a new and interesting method inspired by the ordinary moment estimator. We consider three theoretical statistics: mean, mean of the minimum of any pair of the observations in a mixed distribution, mean

of the minimum of any triplet of the elements in a mixed distribution, and equate them to the observed values. To find the observed values, we simply need to re-sample a data set to sub-samples, first each with sample size of two, and then re-sample it again to sub-samples with three elements each. The parameter estimates can then be found by solving the system of three equations. Our simulation results showed that the performance of this method is plausible, especially for mixtures with medium separation.

All of the four moment based methods discussed above have been compared with the MLE, on their performances in estimating simulated samples drawn from a mixture of two exponential distributions. To investigate the effect of variation of separation between the components and sample sizes, degrees of separation r of 2, 5 and 10 and sample sizes of 10, 15, 20, 50 and 1000 were studied. All methods, even the MLE, perform poorly on small samples. The variance of $\hat{b}$ is particularly large compared to the other two parameter estimates. For large samples, the MLE appears to be the best method in terms of the variance, in spite of the fact that the true values were set as the starting points of the EM algorithm. We have illustrated the sensitiveness of the EM algorithm to the initial values: when initial values deviate by a great extent from the true values, it is quite likely that we will obtain highly biased estimates; the number of iterations may increase, indicating the slow convergence of the EM algorithm. All moment based methods have the ability to provide parameter estimates with high accuracy, given the sample size is large enough. Of course, the difficulties of using the method of moments, for instance the possibility of obtaining estimates with negative value or in complex forms, still apply. It is worth noting that the method of attenuated moments stands out by providing estimates which have lower bias and marginally higher variances compared to the MLE. This method is undoubtedly a good alternative to the MLE as it is quicker and, at the same time, provides parameter estimates with high efficiencies.

9.1.2 Linear Combinations of Exponential Distributions

Consider a Markov process in continuous time on finite state space, if two states are indistinguishable and clumped into a level, it becomes a hidden Markov process. The sojourn time in the level is a linear combination (not necessarily positive mixture) of two exponential distributions. For a time irreversible case, the mixing weight of the first component p can be greater than one. Since the sum of mixing weights is one, the second mixing weight has a negative value. The estimation problem for such a distributions has been a major discussion point in Chapter 5. Since the PDF of a linear combination of two exponential distributions is exactly the same as the one of a positive mixture, all methods studied for an exponential mixture in Chapter 3 can be applied perfectly to the former.

We have stated the conditions for the PDF to be valid, and the ranges of p for the distribution to have a mode. Since the simulation of a data set drawn from a linear combination of exponential distributions is not as straightforward as the positive mixture, we have therefore outlined the procedures for this purpose.

Our simulation experiments showed that the MLE is implausible for the parameter estimation in this case. In most cases, the MLE, as performed by the EM method, fails to identify a linear combination. Instead, it fits a positive mixture to a small sample, and a single exponential distribution to a large sample. There is an issue with the starting values of the EM algorithm: if $p^{(0)}$ is started with a value greater than one, at some point of the iterative process the updated likelihood function will become complex, because $\hat{p}^{(k)}$ has been increased to a value which violates the condition for a valid PDF. After this point, $b^{(k)}$ will decrease until it is close to $a^{(k)}$. This happens regardless of the initial values, hence the MLE always fits a single distribution for a large sample because $\hat{a}$ and $\hat{b}$ have similar values.

The performances of the Appell moment estimator and the method using order statistics have been disappointing; the variances of the estimates b and p are large even when samples are large. However, compared to the MLE, these two methods are much better in recognising a linear combination of two exponential distributions.

Excitingly, the fractional moment estimator and the attenuated moment estimator are able to provide reasonable estimates for a linear combination of two exponential distributions. The attenuated moment estimator is the best method since it provides estimates which are lowest biased and have the smallest variances compared to the other methods.

9.1.3 Mixtures of Geometric Distributions

As a result of Theorem 1 and 2 (see Section 1.7), we understand that the study of almost all properties of N, a random number following a geometric mixture distribution, is analogous to the study of the corresponding properties of T, a random number following an exponential mixture distribution. By replacing the ordinary moments of T with rising factorial moments of N, any method of moments developed for estimating parameters of T can be used to estimate the parameters of N. Therefore, in Chapter 4, we applied the method of rising factorial fractional moments and the method of attenuated rising factorial fractional moments to solve the parameter estimation problem for a mixture of two geometric distributions. Compared to the standard method of rising factorial moments, these two methods perform better by providing estimates with lower bias and variances. We also evaluated the approximated asymptotic variance of estimator for these two methods, which allows us to suggest, respectively, the best fraction, and the optimal combination of fraction and attenuation to guarantee the precision of the estimates given by these two methods. The best fraction in both methods decreases with the separation between the components. In real life, like its continuous analogue, one simply needs to fit a raw sample with a few κ (for the rising factorial fractional moment estimator) or a few combinations of κ and c (for the attenuated moment estimator), and substitute the yielded estimates in the formulae of the asymptotic variances of estimators. The set of parameters that produce the smallest variance should have the highest precision.

A method using double Appell sequences devised by Jalali has been investigated. We have also seen a few examples of Appell double sequences, suggested by Jalali, that can be

used in this method to estimate the parameters of a geometric mixture. In particular, we made use of the Kronecker sequences to fit simulated samples arising from a mixture of two geometric distributions. When the sample size is sufficiently large, this method should be able to provide reasonable estimates of the parameters.

The estimation results of all of the three methods inspired by the method of moments have been compared to the ML approach, based on different sample sizes and various degrees of separation. For small samples, none of the methods stood out to provide plausible estimates. Although the variances of MLEs seemed to be small, the resulting ML inferred distributions have poor fits to the simulated samples of small sizes. On the other hand, the moment based methods are likely to over-estimate the parameter from the second component b when the number of observations in a sample is limited. The reasons for this have been investigated and discussed in Chapter 4.

For samples of large sizes, all these methods provide plausible estimates of the parameters in a mixture of two geometric distributions; the efficiencies of these estimators increase when the components are better separated. A comparison of the performances of all these methods in fitting large samples has been made and, without a surprise, the MLE outperforms other methods in terms of the variance. The rising factorial fractional moment estimator and the Appell moment estimator are not ideal for estimating b when the components are slightly separated. The attenuated moment estimator not only has small variances of estimators which are only marginally larger than the ones given by the MLE, but its estimates also have the lowest bias among all methods when the separation between the components are sufficiently large.

9.1.4 Linear Combinations of Geometric Distributions

If two states are hidden in a level in a Markov chain in discrete time on finite state space, the sojourn time in the level is known to have a linear combination (not necessarily a positive mixture) of two geometric distributions. Like its continuous analogue, a linear combination of geometric distributions has an identical PMF as a positive mixture of geometric distributions. Therefore, any method devised in Chapter 4 can be applied to estimate the parameters in a linear combination of geometric distributions with no difficulty, in spite of the fact that the performances of the method can be affected by the existence of a negative mixing weight. The conditions for the PMF to be valid, and the conditions for the existence of a mode in such a distribution have been outlined in Chapter 5. In the same chapter, we demonstrated how a data set can be simulated from a linear combination of two geometric distributions.

We then illustrated the performances of all four methods: the MLE, the rising factorial fractional moment estimator, the attenuated moment estimator and the Appell moment estimator, in estimating the parameters of a linear combination of two geometric distributions. We have seen that the MLE and the Appell moment estimator have poor performances in this case. The MLE either fits a positive mixture geometric distribution to a small sample,

or it fits a single geometric distribution when the sample size is large ($\hat{a}$ and $\hat{b}$ have similar values). The method based on double Appell sequences provides poor estimates of b: for small samples, $\hat{b}$ are mostly negative which are unrealistic because b as a probability should have a value between 0 and 1; for large samples, the estimates of b are not consistent, as indicated by the large variance of $\hat{b}$.

The rising factorial fractional moment estimator and the attenuated moment estimator are able to provide satisfactory estimates, especially when the number of observations is sufficient. In general, the attenuated moment estimator has the best performance by providing estimates with the minimum bias and variances.

Unlike the positive mixture, we do not find good conformity between the practical and theoretical asymptotic variance of estimator for the method of rising factorial fractional moments and the method of attenuated moments. Despite the approximation error in the evaluation of the theoretical variance and the random error existed in the simulation, the fact that one of the mixing weights is negative has caused the calculation of the theoretical variance to be more complicated. However, the theoretical best fraction and the theoretical optimal combination of κ and c are proven to be able to provide estimates with low variances. Like the positive mixture, the best fraction in both methods reduces in value when the magnitude of separation between the components increases.

9.1.5 Mixture Models for the Incubation Period of Prion Disease

The incubation period of an infectious disease is the time between exposure to an infectious agent and the occurrence of clinical symptoms. It is well known that the incubation periods of most of the infectious diseases are fitted well with a single lognormal distribution. For the prion diseases (such as scrapie, BSE, CJD), the incubation period is a defining characteristic, being very prolonged, and is a key research tool. In Chapter 6, we analysed experimental data consisting of the incubation periods of mice which were exposed to a type of prion disease similar to scrapie. In the experiment, the primary passage mice were exposed, orally, to infectious material derived from a case of chronic wasting disease in deer. When the disease symptoms were manifest, a number of further second passage mice were fed, orally, with the brain of an infected mouse. The experiments were carried on for five passages.

Our earliest attempt to fit full set of the incubation period data with a lognormal distribution had failed to provide satisfactory goodness of fit. Other single distributions, such as exponential, normal, gamma and Weibull, had also been used to fit the data set but, again, none of them was plausible. Since the incubation periods of the earlier passage mice seem to be longer compared to their successors, we undertook two non-parametric tests, the Kolmogorov-Smirnov test and the Mann-Whitney test, and found that the incubation periods of the first passage mice are significantly different from the rest. Therefore, we fitted the overall data set with a mixture of two distributions, in which different component distribution, the lognormal, normal, gamma, Weibull and Burr XII, were used. All mixture

models provide excellent goodness of fit to the real data set.

Out of the five mixture models considered, the gamma mixture model and the lognormal mixture model have lower KS distance. The normal mixture model has the largest log-likelihood function, which is only marginally larger than the others. However, this model is not favoured because it allows negative incubation period, which is not ideal on biological grounds. The Burr XII mixture model has the lowest KS distance, but it has too many parameters to estimate from the small sample which only has 44 observations. Therefore, we prefer the Weibull mixture model, which has a low KS distance, large likelihood values and only a few parameters to estimate.

All of the mixture models also tell a similar story that is relevant to the underlying biology of prion disease: there exists two components in the distribution where the first component has a shorter mean and smaller variance than the second component. Importantly, our observation was confirmed by experimental investigations which found different pathology in infected animals, and thus there are two strains in the system. Prion strains are often characterised by their incubation period patterns, which are stable in a given rodent experimental system. The mixture models therefore provide an ideal framework for identifying different strains in a multiple infection, and to quantify the proportion of each strain at each passage.

9.1.6 Self Revealing Aggregated Markov Processes on Trees (SRAMPT): A Model for Sub-Clinical Infection in Prion Serial Passage Studies

Infectious agents may cause sub-clinical infection, that only manifests as disease on subsequent passage when transmitted to a new host. This is especially the case in the group of prion diseases, where during a pro-longed incubation period, there may be no detectable infection in a host, but the host might still be infectious. In a similar experimental system to the one investigated for incubation periods, the waiting time for the host to exhibit signs of scrapie can be modelled by a special kind of Markov process. With the aim to "track" the sub-clinically infected animals, we showed how a new model constructed by Jalali (2008c), Self Revealing Aggregated Markov Processes on Trees (SRAMPT) can be used to solve this problem and hence give an estimate of the overall prevalence of infection.

We considered three states for an exposed host: "Diseased" (D), "Uninfected and Healthy" (H) and "Sub-clinically Infected" (S). State D is observable, and its successor nodes will always be in this state. States H and S are indistinguishable, but if any successor is found in state D, then we can infer that the predecessors must be S. At a node σ, the infection states are not completely known. We only know the subset $X_\sigma = \{H, S\}$. But, as the process branches out, the information may be revealed on any successor node τ. We named a node with no successor a leaf. With SRAMPT, we write the likelihood

function beyond σ by letting the information percolate from the leaf of each tree, defined as

$$L_{\sigma}^{(i)} = \prod_{\tau \in suc(\sigma)} \sum_{j \in \bar{S}_{\tau}} p_{ij} L_{\tau}^{(j)}.$$

Having found the likelihood function beyond all the root of the tree, the overall likelihood function is then given by

$$\begin{aligned} L_{\langle\rangle} &= \pi_s L_{\langle\rangle}^{(S)} && \text{if } \bar{S}_{\sigma} = \{S\} \\ L_{\langle\rangle} &= (1-\pi_s) L_{\langle\rangle}^{(H)} + \pi_s L_{\langle\rangle}^{(S)} && \text{if } \bar{S}_{\sigma} = \{H, S\}. \end{aligned}$$

We maximise $L_{\langle\rangle}$ to obtain the estimates of p_{sd}, p_{sh} and π_s.

We analysed experimental data provided by Professor A. Aguzzi, Institute of Neuropathology, University of Zurich. A naïve estimate of the prevalence of scrapie at first passage is 17%, and without a model framework we are unable to draw conclusions about sub-clinical transmission rates. We applied SRAMPT and found that the prevalence of infection at first passage was much higher than assumed, at 26%. The SRAMPT model provided further biological insights, that there is a 22% chance for the "contact" of a sub-clinically infected mouse to develop scrapie disease. Sub-clinically infected hosts are almost certain to pass on infection, although most will remain sub-clinical ($\approx$ 74%). We conclude that SRAMPT provides a promising means for tracking sub-clinical effects that cannot be directly observed, and may be applicable in a wide range of experimental and epidemiological systems.

9.1.7 Application of the SRAMPT Model to Epidemiological Contact Chains

The SRAMPT model was initially developed for a particular system: the serial passage prion disease experiments which generated the incubation period data analysed in Chapter 6. However, it soon became apparent that with minor modifications, the model could be used to investigate transmission of sub-clinical infection in chains of epidemic contacts of a range of other infectious diseases. We make only one change to the Markov process, such that D is no longer an absorbing state. We found that this greatly complicated the parameter estimation process. However, using an approach based on Jalali (2008c), which estimates the transition probabilities by maximising the likelihood function of the Markov model, we found solutions, or a range of solutions, for all parameters. For such a model, a chain of states can be viewed as two parts: (1) tail which starts with D and ends with D, and consists of a sequence of D's and S's (2) head which contains k E's. (note that E is the sate which we cannot tell if a node is in state H or S). Since the number of transitions from D to self-revealed S must be identical to the number of transitions from self-revealed D, we let x be the product of p_{ds} and p_{sd}. This reduces the number of parameters to be estimated,

but it also means that we will only get a range of estimates for some of the parameters. The ML equations can then be solved using Kuhn-Tucker relations to obtain the parameter estimates. To summarise, p_{dd} is always obtainable in an exact form, p_{ss} is estimated by solving the *Fundamental Equation*, x can then be found with the knowledge of k, n_0, n_1 and p_{ss}. With these three parameter estimates, we then find the extreme solutions of p_{ds}, p_{sd}, p_{dh} and p_{sh}. It is worth mentioning that the Markov model can be reduced into two states if the *Fundamental Inequality* does not hold.

We tested the model on a set of simulations designed to reflect the type of epidemiological data collected during contact tracing exercises in disease outbreaks. The epidemic chain represents a set of individuals linked by contact and potentially exposed to disease via an index case, in state D, at some point earlier in the chain.

We explored the typical disease scenario, where, although sub-clinical infection can occur, most disease transmission occurs from clinically positive (D) individuals. The success of the parameter estimation was dependent on sample size, and we found that approximately 20 chains were required for 'high' transmission rates and 50 - 150 were required for 'low' transmission rates. We note that in this system of hidden states, some parameters could not be uniquely identified. However, we found that by making the simple assumption that sub-clinical infections were more likely to fail to transfer infection than clinical cases, all parameters could be identified with an impressive precision.

9.2 Future Research

9.2.1 Mixtures and Linear Combinations of Distributions

All moment based methods discussed in this thesis are quick and simple approaches to solve the estimation problem of a linear combination (both the positive mixture and the case when $p > 1$) of two exponential/geometric distributions. For a sufficiently large sample arising from a positive mixture distribution, they are able to provide good estimates of parameters, and their performances are comparable to the desirable maximum likelihood estimator. Since the MLE is known to be sensitive to the starting points in the EM algorithm, the easy methods investigated in this thesis will act as good tools to provide good starting values for the MLE. Recommended future studies include the extension of the approaches presented here to other types of components, e.g. Weibull, normal, binomial etc., or to mixture distributions with more components ($m \geq 3$). A mixture model with more components should be better in capturing specific properties of real data. The price paid for this flexibility is that the amount of algebra involved in the estimation problem is even larger than the ones we studied in this thesis. By increasing m by one, the model will have two more parameters to deal with. Therefore, we should expect that, in order to obtain good parameter estimates, the amount of data available should be very large.

For a linear combination of distributions, at least one of the mixing weights can be

negative. In real life, for the data coming especially from clumped Markov processes, there may be some negative weight present. As the PDF of such a distribution, in algebraic form, is identical to a mixture distribution, it will be useful if we have a test which distinguishes a linear combination from a positive mixture. Indeed, Jalali (2009) has recently constructed some tests for this purpose. In the future, we will study the power of his tests based on some simulation experiments.

Since we only found two good methods (the fractional moment estimator and the attenuated moment estimator) for the parameter estimation of a linear combination of two distributions, a priority in future research will be to investigate the performance of other methods, for example the Bayesian approach or the minimum distance estimator, in such a problem.

Throughout this thesis, we have compared the performances of the moment based methods with the asymptotically most efficient MLE in the estimation problem of a linear combination of two exponential distributions, and the discrete geometric analogue. Future researchers may compare these methods with the well known Bayesian approaches, where the prior distributions of the parameters have to be specified.

The literature surrounding the problem of assessing the number of components in mixture models is large because the problem is important but very difficult. Since this problem has not been completely resolved, further research is needed so that the number of components in a mixture distribution is better determined. For example, in Section 3.6.1, we have learned that, according to Jalali's (2007) paper, $\eta > \phi > \tau$ and $\eta\tau - \phi^2 > 0$ are necessary for a proper mixture of exponentials. Therefore, one can investigate the distribution of $\eta\tau - \phi^2$ for a mixture of exponential distributions and hence provide a test to identify if a real data comes from a mixture of two exponential distributions.

9.2.2 Incubation Period Models, SRAMPT and Sub-Clinical Infections

The primary applications of incubation periods in prion research are either to quantify dose, or to identify strains. The latter is particularly important for classifying strains of unknown origin against known strains (for example the classification of early vCJD cases with BSE). This classification is often performed informally, and there is a need for a robust statistical method for characterising and comparing prion incubation periods. The mixture model potentially provides such a framework. There exists large archives of studies of scrapie strains, that could be used to further test the model (Maclean & Bostock (2000)), and investigate the effects of dose (which tend to prolong the incubation period).

The SRAMPT model initially had a specific purpose: to estimate the prevalence of sub-clinical infection in a serial passage study of scrapie. The output of the analysis has proved very useful in highlighting considerable disease transfer in a system of de novo generated prions, that was not apparent in a simple consideration of the data. The next application of the model will be to estimate the sub-clinical prevalence in similar experimental systems that need comparison with a control group. In this case it will be important to investigate whether

the prevalence is above a threshold expected from exposures under control conditions.

For the application of SRAMPT to epidemic contact chains, the next step is clearly to apply the method to a real data set. There are a number of possibilities. Probably the most sensible starting point would be tuberculosis, for which sub-clinical infection is common and for which the asymptomatic state is known to be infectious. It is assumed that the level of infectiousness is much less than for a clinical case, but this could be quantified (for the first time) using SRAMPT. Contact tracing is performed during all UK tuberculosis outbreaks, and data sets are potentially available. Sub-clinical infection is also a possible route of transfer for influenza virus. Large contact chains have been collected during recent outbreaks, and the model could be applied to help detect and quantify any sub-clinical effects.

In the future application of the models developed in this thesis, we can identify two broad ambitious goals.

1. **A general statistical model for prion serial passage study.**

   This will require the integration of the work in Chapters 6 and 7. We note that in the serial passage study we have only modelled the presence or absence of infection. The changes in the incubation period, in each passage, could also be considered.

2. **A general statistical model for sub-clinical transmission in epidemiological networks.**

   A contact tracing exercise during a disease outbreak does not always result in complete, distinct, chains of infection. Incomplete data is common, and epidemiological networks can be complex. To analyse these situations will require the extension of the model in Chapter 8 to encompass, first, branching, and finally clusters of contacts that cannot be represented by simple trees.

The work here represents a start to both these projects, and shows how well chosen statistical models can shed important new light onto biological systems.

Bibliography

Aitkin, M., Anderson, D. & Hinde, J. (1985). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statisitical Society Series A* **144**, 419 – 461.

Albert, J. R. G. & Baxter, L. A. (1995). Applications of the EM algorithm to the analysis of life length data. *Applied Statistics* **44**, 323 – 341.

Anderson, R. M., Donnelly, C. A. & et al (1996). Transmission dynamics and epidemiology of BSE in British castle. *Nature* **382**, 779 – 788.

Armenian, H. K. & Khoury, M. J. (1981). Age at onset of genetic diseases. *American Journal of Epidemiology* **113**, 596 – 605.

Armenian, H. K. & Lilienfeld, A. M. (1974). The distribution of incubation periods of neoplastic diseases. *American Journal of Epidemiology* **99**, 92 – 100.

Barria, M. A., Mukherjee, A., Gonzalez-Romero, D., Morales, R. & Soto, C. (2009). De novo generation of infectious prions in vitro produces a new disease phenotype. *PLoS Pathogens* **5 (5): e1000421**.

Bartholomew, D. J. (1969). Sufficient conditions for a mixture of exponentials to be a probability density function. *The Annals of Mathematical Statistics* **40**, 2183 – 2188.

Behboodian, J. (1970). On a mixture of normal distributions. *Biometrika* **57**, 215 – 217.

Bening, V. E., Korolev, V. Y., Kolokol'tsov, V. N., Saenko, V. V., Uchaikin, V. V. & Zolotarev, V. M. (2004). Estimation of parameters of fractional stable distributions. *Journal of Mathematical Sciences* **123**, 3722 – 3732.

Besbeas, P. & Morgan, B. J. T. (2004). Efficient and robust estimation for the one-sided stable distribution of index 1/2. *Statistics Probability Letters* **66**, 251 – 257.

Bhattacharya, C. G. (1967). A simple method of resolution of a distribution into Gaussian components. *Biometrics* **23**, 115 – 135.

Bignami, A. & de Matteis, A. (1971). A note on sampling from combinations of distributions. *Journal of the Institute of Mathematics and its Applications* **8**, 80 – 81.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31 – 38.

Blischke, W. R. (1962). Moment estimators for the parameters of a mixture of two binomial distributions. *The Annals of Mathematical Statistics* **33**, 444 – 454.

Blischke, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association* **59**, 510 – 528.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. & Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* **46**, 373 – 388.

Bruce, M. E. (1993). Scrapie strain variation and mutation. *British Medical Bulletin* **49**, 822 – 838.

Bruce, M. E., Will, R. G., Ironside, J. W. & et al (1997). Transmissions to mice indicate that 'new variant' CJD is caused by the BSE agent. *Nature* **389**, 498 – 501.

Burr, I. W. (1942). Cumulative frequency function. *Annals of Mathematical Statistics* **13**, 215 – 232.

Cassie, R. M. (1954). Some uses of probability paper in the analysis of size frequency distributions. *Australian Journal of Marine Freshwater Research* **5**, 513 – 522.

Charlier, C. V. L. (1906). Researchers into the theory of probability. *Lunds Universitets Årskrift, Ny följd* **2.1, No. 5**.

Charlier, C. V. L. & Wicksell, S. D. (1924). On the dissection of frequency functions. *Arkiv för Matematik, Astronomi och Fysik* **18**, 1 – 64.

Choi, S. C. & Wette, R. (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics* **11**, 683 – 690.

Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics* **9**, 15 – 28.

Cox, D. R. (1966). Notes on the analysis of mixed frequency distributions. *British Journal of Mathematical and Statistical Psychology* **19**, 39 – 47.

Davis, D. J. (1952). An analysis of some failure data. *Journal of the American Statistical Association* **47**, 113 – 150.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463 – 474.

Deely, J. J. & Kruse, R. L. (1968). Construction of sequences estimating the mixing distribution. *The Annals of Mathematical Statistics* **39**, 286 – 288.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* **B 39**, 1 – 38.

Detwiler, L. A. (1992). Scrapie. *Revue Scientifique et Technique, International Office of Epizootics* **11 (2)**, 491 – 537.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 363 – 375.

Eisenberger, I. (1964). Genesis of bimodal distributions. *Technometrics* **6**, 357 – 363.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *M. D. Escobar and M. West* **90**, 577 – 588.

Everitt, B. S. & Hand, D. J. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

Falls, L. W. (1970). Estimation of parameters in compound Weibull distributions. *Technometrics* **12**, 399 – 407.

Fowlkes, E. B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association* **74**, 561 – 575.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Fryer, J. G. & Robertson, C. A. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika* **59**, 639 – 648.

Gajdusek, D. C., Gibbs, C. J. & Alpers, M. P. (1966). Experimental transmission of a kuru-like syndrome to chimpanzees. *Nature* **209**, 794 – 796.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398 – 409.

Godoment, R. (1969). *Algebra*. Kershaw Publications.

Gravenor, M. B. (2003). Analysis of serial passage PrP experimental data 2, unpublished .

Gravenor, M. B., Stallard, N., Curnow, R. & McLean, A. R. (2003). Repeated challenge with prion disease: The risk of infection and impact on incubation period. *Proceedings of the National Academy of Sciences USA* **100**, 10,960 – 10,965.

Gruet, M., Philippe, A. & Robert, C. P. (1999). MCMC control spreadsheets for exponential mixture estimation. *Journal of Computational and Graphical Statistics* **8**, 298 – 317.

Hadlow, W. J. (1959). Scrapie and kuru. *The Lancet* **ii**, 289 – 290.

Haldane, J. B. S. (1952). Simple tests for bimodality and bitangentiality. *Annals of Eugenics* **16**, 359 – 364.

Harding, J. P. (1949). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association of the UK* **28**, 141 – 153.

Harris, D. (1968). A method of separating two superimposed normal distributions using arithmetic probability paper. *The Journal of Animal Ecology* **37**, 315 – 319.

Harris, H. & Smith, C. A. B. (1949). The sib-sib age of onset correlation among individuals suffering from the same heredity syndrome produced by more than one gene. *Annals of Eugenics* **14**, 309 – 318.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics* **8**, 431 – 446.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association* **64**, 1459 – 1471.

Hill, A. F. (2000). Species-barrier-independent prion replication in apparently resistant species. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10,248 – 10,253.

Hill, A. F. & Collinge, J. (2003). Subclinical prion infection. *Trends in Microbiology* **11**, 578 – 584.

Horner, R. D. (1987). Age at onset of Alzheimer's disease: Clue to the relative importance of etiologic factors? *American Journal of Epidemiology* **126**, 409 – 414.

Hosmer, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics* **29**, 761 – 770.

Jalali, A. (2002). Two classes of mixtures II: On mixtures of Polya-Laguerre frequency functions of finite class, unpublished .

Jalali, A. (2005a). A note on the discrete version of estimation by Appell sequences, unpublished .

Jalali, A. (2005b). On a method of parameter estimation for mixtures of exponentials based on an Appell sequence, unpublished .

Jalali, A. (2005c). On the discrete version of mixtures of Polya-Lagguer finite class, unpublished .

Jalali, A. (2006). A general method to calculate the asymptotic covariance matrix of generalised moment estimator, unpublished .

Jalali, A. (2007). Estimation of parameters of mixtures of exponential distributions using order statistics, unpublished .

Jalali, A. (2008a). Differential geometry of mixture of two exponentials part i: The Fisher information matrix, unpublished .

Jalali, A. (2008b). On mixtures of distributions from an exponential family part i: Single parameter families, unpublished .

Jalali, A. (2008c). On self-revealing aggregated markov processes on trees (SRAMPT), unpublished .

Jalali, A. (2009). Test for the positiveness of linear combinations of exponential and geometric distributions, unpublished .

Joffe, A. D. (1964). Mixed exponential estimation by the method of half moments. *Applied Statistics* **13**, 91 – 98.

John, S. (1970). On identifying the population of origin of each observation in a mixture of observations from two gamma populations. *Technometrics* **12**, 565 – 568.

Kao, J. H. K. (1959). A graphical estimation of mixed Weibull parameters in life-testing electron tubes. *Technometrics* **1**, 389 – 407.

Karlis, D. & Xekalaki, E. (1998). Minimum Hellinger distance estimation for finite Poisson mixtures. *Computational Statistics and Data Analysis* **29**, 81 – 103.

Kondo, K. (1977). The lognormal distribution of the incubation time of exogenous diseases. genetic interpretations and a computer simulation. *Japanese Journal of Human Genetics* **21**, 217 – 237.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79 – 86.

Laird, N., Lange, N. & Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association* **82**, 97 – 105.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics* **22**, 1081 – 1114.

Lindstrom, M. J. & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**, 1014 – 1022.

Lo, Y., Mendell, N. R. & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* **88**, 767 – 778.

Maclean, A. R. & Bostock, C. (2000). Scrapie infections initiated at varying doses: An analysis of 117 titration experiments. *Philosophical Transactions of the the Royal Society of London Series B* **355**, 1043 – 1050.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318 – 324.

McLachlan, G. J. R. & Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley.

Mendenhall, W. & Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika* **45**, 504 – 520.

Ng, S. K. & McLachlan, G. J. (1998). On modifications to the long-term survival mixture model in the presence of competing risks. *Journal of Statistical Computation and Simulation* **61**, 77 – 96.

Pattison, I. H. (1965). *Experiments with Scrapie with Special Reference to the Nature of the Agent and the Pathology of Disease. In Slow, Latent and Temperature Virus Infections (NINDB Monograpgh No. 2) (Gajdusek, D. C. Et. Al., Eds)*. US Government Printing Office.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185**, 71 – 110.

Philippe, P. (1990). Twinning causative origin investigated by Sartwell's biometrical method. *American Journal of Human Biology* **2**, 107 – 115.

Prusiner, S. B. (1982). Novel proteinaceous infectious particles cause scrapie. *Science* **216**, 136 – 144.

Prusiner, S. B., Cochran, S. P., Groth, D. F., Deborah, A. B., Downey, E., Bowman, K. A. & Marinez, H. M. (1981). Measurement of the scrapie agent using an incubation time interval assay. *Annals of Neurology* **11**, 353 – 358.

Race, R., Meade-White, K., Raines, A., Raymond, G. J., Caughey, B. & Chesebro, B. (2002). Subclinical scrapie infection in a resistant species: Persistence, replication, and adaptation of infectivity during four passages. *The Journal of Infectious Diseases* **186**, S166 – S170.

Race, R., Raines, A., Raymond, G. J., Caughey, B. & Chesebro, B. (2001). Long-term subclinical carrier state precedes scrapie replication and adaptation in a resistant species:

Analogies to Bovine Spongiform Encephalopathy and variant Creutzfeldt-Jakob disease in humans. *Journal of Virology* **75**, 10,106 – 10,112.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *Annals of Statistics* **26**, 195 – 239.

Rider, P. R. (1961). The method of moments applied to a mixture of two exponential distributions. *The Annals of Mathematical Statistics* **32**, 143 – 147.

Sartwell, P. E. (1950). The distribution of incubation periods of infectious disease. *American Journal of Hygiene* **51**, 310 – 318.

Seidel, W., Mosler, K. & Alker, M. (2000a). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics* **52**, 481 – 487.

Seidel, W., Mosler, K. & Alker, M. (2000b). Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models. *Statistical Papers* **41**, 85 – 98.

Slater, L. J. (1966). *Generalized Hypergeometric Functions*. Cambridge, England: Cambridge University Press.

Tallis, G. M. & Light, R. (1968). The use of fractional moments for estimating the parameters of a mixed exponential distribution. *Technometrics* **10**, 161 – 175.

Tan, W. Y. & Chang, W. C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association* **67**, 702 – 708.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528 – 540.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Ueda, N. & Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks* **11**, 271 – 282.

Vounatsou, P., Smith, T. & Smith, A. F. M. (1998). Bayesian analysis of two-component mixture distributions applied to estimating Malaria attributable fractions. *Applied Statistics* **47**, 575 –587.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* **57**, 307 – 333.

Watkins, A. J. (1999). An algorithm for maximum likelihood estimation in the three parameter Bur XII distribution. *Computational Statistic Data Analysis* **32**, 19 – 27.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics* **18**, 293 – 297.

Whittle, H. C., Aaby, P., Samb, B., Jensen, H., Bennett, J. & Simondon, F. (1999). The effect of subclinical infection on maintaining immunity againsts measles in vaccinated children in West Africa. *The Lancet* **353**, 98 – 101.

Wilesmith, J. W., Wells, G. A., Cranwell, M. P. & Ryan, J. B. (1988). Bovine Spongiform Encephalopathy: Epidemiological studies. *The Veterinary Record* **123**, 638 – 644.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9**, 60 – 62.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* **5**, 329 – 350.

Woodward, W. A., Parr, W. C., Schucany, W. R. & Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *Journal of the American Statistical Association* **79**, 590 – 598.

Yao, Q. & Morgan, B. J. T. (1999). Empirical transform estimation for indexed stochastic models. *The Journal of the Royal Statistical Society Series B* **61**, 127 – 141.